SoK: On the Offensive Potential of AI

Paper accepted at the 3rd IEEE Conference on Secure and Trustworthy Machine Learning (SaTML'25)

Saskia Laura Schröer[*], Giovanni Apruzzese[*], Soheil Human[†], Pavel Laskov[*],
Hyrum S. Anderson[‡], Edward W. N. Bernroider[†], Aurore Fass[§], Ben Nassi[¶], Vera Rimmer[‖], Fabio Roli[**],
Samer Salam[††], Ashley Shen[††], Ali Sunyaev[‡‡], Tim Wadhwa-Brown[††], Isabel Wagner[x], Gang Wang[xi]
[*]*University of Liechtenstein*, [†]*Vienna University of Economics and Business*, [‡]*Robust Intelligence*, [§]*CISPA*, [¶]*Technion*,
[‖]*KU Leuven*, [**]*University of Genoa*, [††]*CISCO Systems*, [‡‡]*Karlsruhe Institute of Technology*, [x]*University of Basel*, [xi]*UIUC*
Corresponding authors: saskia.schroeer@uni.li, giovanni.apruzzese@uni.li, soheil.human@wu.ac.at, pavel.laskov@uni.li

***Abstract*—Our society increasingly benefits from Artificial Intelligence (AI). Unfortunately, more and more evidence shows that AI is also used for offensive purposes. Prior works have revealed various examples of use cases in which the deployment of AI can lead to violation of security and privacy objectives. No extant work, however, has been able to draw a holistic picture of the offensive potential of AI. In this SoK paper we seek to lay the ground for a systematic analysis of the heterogeneous capabilities of offensive AI. In particular we (i) account for AI risks to both humans and systems while (ii) consolidating and distilling knowledge from academic literature, expert opinions, industrial venues, as well as laypeople—all of which being valuable sources of information on offensive AI.**

To enable alignment of such diverse sources of knowledge, we devise a common set of criteria reflecting essential technological factors related to offensive AI. With the help of such criteria, we systematically analyze: 95 research papers; 38 InfoSec briefings (from, e.g., BlackHat); the responses of a user study (N=549) entailing individuals with diverse backgrounds and expertise; and the opinion of 12 experts. Our contributions not only reveal concerning ways (some of which overlooked by prior work) in which AI can be offensively used *today*, but also represent a foothold to address this threat in the *years to come*.

***Index Terms*—cyber security, machine learning, society**

I. Introduction

Artificial Intelligence (AI) is an exemplary use-case of a disruptive technology [1, 2]. AI has revolutionized the IT ecosystem worldwide, providing cost-effective solutions for new and existing tasks—potentially exceeding the proficiency of humans [3–5]. Unfortunately, the disruptive nature of AI also has gradually materialized in a more literal sense—as a means to *realize, facilitate and enhance cyberattacks*. Such an observation underscores that the potential of AI must be proactively scrutinized from a cybersecurity perspective.

The domains of AI and cybersecurity are, in fact, strongly intertwined. Abundant works highlight the potential of "AI for cybersecurity" [6], e.g., showing that AI can improve cybersecurity routines [7, 8]; or that AI can perform tasks otherwise unfeasible for security operators [9]. At the same time, a large body of literature focuses on "security of AI" [10, 11], e.g., elucidating that AI methods can be broken with tiny perturbations [12]; or that some confidential information pertaining to AI solutions (i.e., training data, or the AI model itself) can be leaked [13] or stolen [14]. There is another use case, however, that links AI and cybersecurity, but has not received the same degree of attention so far: "offensive AI."

Some prior works have considered scenarios wherein AI is used as an offensive tool. For instance, using Large Language Models (LLM) to write phishing emails [15] is cheap and effective [16], and evidence shows that this is already happening [17, 18]. However, no prior work has systematically analyzed the topic of offensive AI, examining a broad range of attack targets and accounting for diverse sources of knowledge. Indeed, prior systematizations (e.g., [19, 20]) mostly accounted for the viewpoint of academic literature, which is a profound but not the only source of information. Briefings of *industrial conferences*, opinions of *experts*, and even *laypeople* provide complementary perspectives on the ins-and-outs of offensive AI. Furthermore, the offensive potential of AI poses a threat not only to IT *systems* in a narrow sense, as primarily considered in prior work, but also to any stakeholder relying on them, e.g., *humans*, or even the *society* as a whole. Hence, to tackle today's unforeseen risks of offensive AI, a broader scope must be considered for systematization of knowledge, as suggested schematically in Fig. 1. New knowledge sources and versatile use-cases should be taken into account for a comprehensive analysis of this inescapable threat.

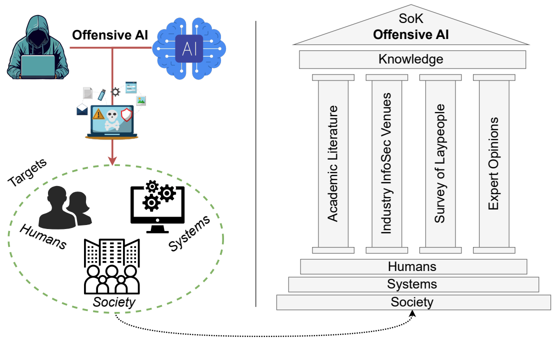


Fig. 1: **Targets and knowledge sources related to *offensive AI*.**

In this SoK, we seek to establish a foundation for understanding and mitigating the (current and future) offensive potential of AI. To this end, we make three high-level **contributions**:

ℂ1: We present a **snapshot of the current landscape** of offensive AI, by accounting for its three crucial stakeholders: systems, humans, and society. This contribution serves to *review its various use cases*, including those potentially overlooked by prior surveys on offensive AI.

ℂ2: We devise an original **long-term classification of key**

technological factors related to offensive AI. This contribution serves to examine and compare selected works on offensive AI according to a *common set of criteria*. We develop an online tool [21] to make this contribution applicable to any relevant works (potentially omitted in our literature analysis), including the future ones.

ℂ3: We outline an actionable research agenda for future work, pinpointing **open problems and concerns to be addressed by the research community** in various areas. This contribution serves as a *guide for any stakeholder* interested in mitigating the threat of offensive AI.

Organization. We define the scope of our SoK and describe the methods used to carry out our systematization in Section §II. There, we also present an original *checklist* providing the groundwork for our analyses (and also for ℂ2).

Section §III is devoted to the analysis of *academic literature*. We scrutinize 95 peer-reviewed papers on offensive AI (identified through a systematic search across more than 3000 works) under the lens of our checklist. Our findings reveal several gaps in prior work; for instance, that certain offensive AI use-cases targeting humans, e.g., attribute inference attacks [22, 23], were left out in prior literature reviews.

In Section §IV, we focus on the *industrial perspective*. We survey the landscape of major InfoSec outlets (BlackHat, DefCon) and identify 38 briefings related to offensive AI. We systematize these works through our checklist, underscoring the offensive potential of AI revealed in practical venues.

In Section §V, we consider the *viewpoint of laypeople*. We present the results of a user study (n=549) exposing the perception of offensive AI by "non-experts." We observe that a large share of participants (84%) are concerned about the potential offensive use of AI and qualitatively analyze the reasons for such concerns—revealing some potential misconceptions.

Finally, in Section §VI, we study the *opinion of experts* on the offensive potential of AI. We reached out to 12 experts in cybersecurity, privacy, and information systems. First, we asked them to complete the questionnaire we used for the general public study. Then, we provided these experts with an early draft of this paper, and requested to write statements defining "three open problems in the field of offensive AI." We systematically review and coalesce these statements (reported verbatim in Appendix D) into ten open problems and fundamental concerns of offensive AI—the basis of ℂ3.

We wrap up our systematization in Section §VII by reflecting on the takeaways on *all hitherto analyzed sources of knowledge* (which collectively form the snapshot of ℂ1, set up the stage for ℂ2, and are used for ℂ3). We also identify and discuss limitations, and compare our contributions with previous related work. We conclude our SoK in Section §VIII.

Scope. In our SoK, we consider offensive AI (OAI) as *the means of using AI to accomplish a task that violates security and privacy objectives*. Such a broad notion covers a wide array of risks, stemming from an attacker who is deliberately trying to cause harm—and does not cover cases in which, e.g., an AI leads to harm due to negligence or misconfiguration. Specifically, our notion encompasses cases when AI is used to amplify existing threats (e.g., disinformation is a well-known problem which can be made much worse via AI [24]) or develop previously unseen threats (e.g., attribute inference attacks are essentially enabled by AI [22]).[a]

[a] **Prior work.** We summarize the evolution of the term "offensive AI" in the literature in Appendix E-B. Some works associate techniques for generation of "adversarial examples" to OAI [20]. According to our definition, *some instances* of such techniques can be considered as OAI (e.g., if an attack involves Generative Adversarial Networks [25, 26], which clearly rely on AI), while others are orthogonal to OAI (e.g., some "evasion" attacks [27] not necessarily rely on AI to be staged—i.e., computing FGSM to generate an adversarial perturbation can be done algorithmically without leveraging any AI technique [10]). We stress that we use "AI" to denote techniques within the machine-learning (ML) domain [28].

II. Research Methods and Checklist

We introduce the research methods applied in our SoK: the systematization of scientific literature (§II-A) and of InfoSec briefings (§II-B), the user study with non-experts (§II-C), and the elicitation and systematization of expert knowledge (§II-D). We also present our *OAI Assessment Checklist*, which provides the means for alignment and systematization of diverse classes of prior work considered in our SoK (§II-E). Some details of our methods are in the Appendix, including a timeline (in Fig. 17) encompassing all our research activities.

A. Systematic Literature Review (Methodology)

Prior surveys on OAI [20, 29, 30] are grounded in academic literature. Hence, to ensure continuity, we consider research papers as our first source of knowledge. We perform a *systematic literature review*, following established guidelines [31], illustrated in Fig. 2. We describe the pivotal points below.

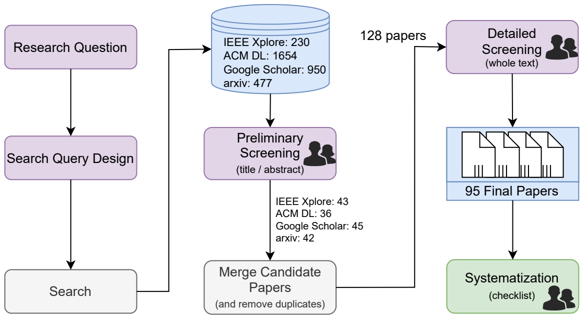


Fig. 2: **Systematic Literature Review.** We collect over 3000 papers from various repositories. After filtering, screening and inter-researcher discussions, we coalesce 95 papers on OAI which we consider in this SoK.

Search Queries. We began our literature review by asking ourselves "how has prior work envisioned *offensive AI*?" To systematically encompass a broad spectrum of prior art, we search for related papers indexed by four popular databases (until Nov. 2023): IEEE Xplore, Google Scholar, ACM DL, and arXiv (similarly to Ladisa et al. [32]). We carried out our search by formulating queries corresponding to two macro-search queries. Specifically, the first, straightforward, macro search-query entails "offensive AI." However, we are confident that there are other ways in which prior work has conceived AI-based applications that can fall within our definition of OAI. Hence, as an exemplary use-case to extend our search, we consider another macro-query revolving around "AI in offensive security": indeed, the idea of using AI in penetration testing (e.g., [33, 34]) *can* also be leveraged by real attackers to bypass a given system. We have even reached out to the authors of some well-known publications (e.g., [35–37]) who confirmed that their AI-based tool could be used maliciously. Overall, based on the our two macro-search queries, we devise 36 search queries, condensed in the box below:

> *Macro query #1*: "offensive" $\land$ ("AI" $\lor$ "artificial intelligence" $\lor$ "ML" $\lor$ "machine learning") $\land$ ("security" $\lor$ "cyber security" $\lor$ "cybersecurity" $\lor$ "cyber attack" $\lor$ "attack" $\lor$ "privacy" $\lor$ "threat");
> *Macro query #2*: ("AI" $\lor$ "artificial intelligence" $\lor$ "ML" $\lor$ "machine learning") $\land$ ("penetration testing" $\lor$ "red teaming").

Importantly, we are aware that our search strategies have limitations and cannot capture "all" works that have considered OAI applications. This is, however, not our goal. We elaborate on how to overcome this limitation in §II-E and §VII.

Dual Reviewing (with adjudication). Our search returned 3311 papers. To mitigate bias, these papers have been reviewed by two authors who worked independently and later compared their findings to find a consensus; in cases of disagreement, a senior reviewer acted as adjudicator [38]. Such a system was used for two steps of our literature analysis:

- *Screening.* First, we identified unique works that fall in our definition of OAI. This was done mostly by inspecting the title and abstract of the papers, which was typically sufficient to remove papers outside our scope; if we were uncertain, we also looked at the entire content of the papers.[1] After removing duplicates, we derived a set of 95 papers.
- *Systematization.* We systematically assess these 95 works. First, we differentiate "technical" from "non-technical" papers: technical papers *must* demonstrate a practical implementation/usage of an AI model; in contrast, non-technical papers encompass case studies, user/expert surveys, conceptual papers, opinion papers, or similar. This classification yielded 16 non-technical and 79 technical papers. Then, we systematize these works according to our checklist (described in §II-E), and we further scrutinize technical papers to underscore technical aspects of their implementation.

B. Systematic Analysis of InfoSec Briefings (Methodology)

Since the emergence of ChatGPT, OAI has become a focal point of discussion, often featured in the news such as by Forbes, Economist, or CNN [44–46]. While non-academic literature can include different types of works, such as news articles, or security reports, our objective is to *also* scrutinize prior work that has not undergone an academic publishing process, but that *(i)* still allows us to identify OAI use cases, *(ii)* is highly relevant to practical and real-world threats, and *(iii)* has been subject to some kind of review process. Hence, we consider the content of two renown security events: BlackHat and DefCon. These venues are highly competitive: for instance, the acceptance rate for the AI track of BlackHat Asia'24 was 7% [47]. To the best of our knowledge, this is the first SoK to consider the perspective of InfoSec "briefings," i.e., presentations of 30–40m with slides and abstract.

We examined the entire history of these venues, from 1993 to 2023 for DefCon, and from 1997 to 2023 for BlackHat. For BlackHat, we assessed the content of the events held in the USA (27), Europe (23), and Asia (19), for a total of 69 events. For DefCon, we assessed all 31 events. We then followed a similar procedure for our literature analysis, rooted in the dual reviewing with adjudication system. First, we looked at all the briefings trying to identify which were related to OAI. To this end, we first inspected the title and abstract; then we performed a deeper analysis by going through the slides, the video, and even the captions of the recording (if available; we could not find any of these resources for [48] which we exclude). Ultimately, we identified 38 briefings related to OAI, which we scrutinize according to our checklist (§II-E).

> We show in Fig. 15 (Appendix E) the yearly distribution of the works (95 papers and 38 briefings) considered in our SoK. Intriguingly, the earliest work for each category appeared in 2008 [49, 50]

C. Study of "non-expert" Opinion (Methodology)

Literature and briefings provide extensive knowledge on OAI; yet, they may not capture what OAI-related concerns are predominant in the real world. To provide a complementary perspective that allows one to ascertain more transient forms of knowledge (as also done in other SoKs, e.g., [51, 52]), we also investigate the perception of OAI among individuals who are not necessarily subject-matter experts.

Questionnaire Design. We devise an anonymous questionnaire covering various aspects related to OAI. Our questionnaire is *short* ($\sim$5 minutes according to five pilot tests) to maximize the response rate and improve the quality of responses. After informing our participants of their rights and collecting some (optional) demographic details (we do not, e.g., ask for specific employment information), we ask up to four questions, visualized in Fig. 3. Potentially, the questionnaire may end after just the first question—which serves as a "screening," so that only participants who have thought about OAI are requested to elaborate their concerns/ideas. Our questionnaire[2] is provided (verbatim) in our repository [54].

Dissemination. To ensure a diverse respondent pool, we distributed the survey over various channels, spanning across online social networks (e.g., one author made posts on LinkedIn—ensuring not to mention "OAI" anywhere in the text) as well as educational events (e.g., lectures and workshops—not related to OAI), and we also relied on convenience sampling [55, 56]. Importantly, we *never primed* our participants: the events were not about OAI, the posts used to share our questionnaire did not link to external resources

[1] We omitted: *(i)* papers that do not present any OAI capability—e.g., AI for security (e.g. [39]), or security of AI (e.g. [40]), or unrelated to security (e.g., [41]); *(ii)* literature reviews on a specific sub-field, e.g., AI in penetration testing [42]; and *(iii)* grey literature/white papers [43].

[2] **Ethical Statement:** we treated our participants ethically, following the Menlo report [53]. We did not ask for personally identifiable information, and our participants can ask us to delete their data if they so desire. Our institutions are aware of our research. Participation in our questionnaire was voluntary and we did not offer any form of compensation.

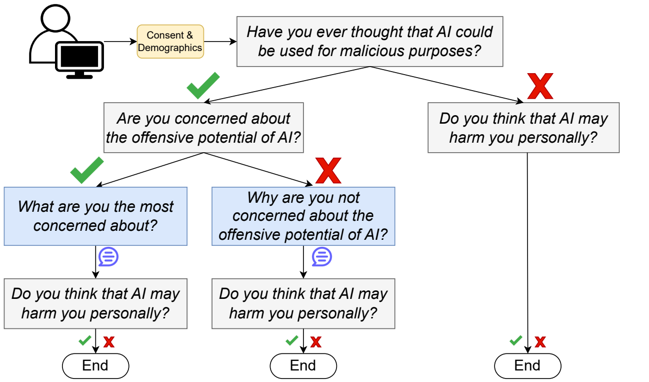

Consent & Demographics
Have you ever thought that AI could be used for malicious purposes?
Are you concerned about the offensive potential of AI?
Do you think that AI may harm you personally?
What are you the most concerned about?
Why are you not concerned about the offensive potential of AI?
Do you think that AI may harm you personally?
Do you think that AI may harm you personally?
End
End
End

Fig. 3: **Questionnaire.** Depending on the answers, participants have to respond to up to four broad questions (e.g., no specific time frame is given). Some questions expect open answers, ensuring freedom to share any concern.

related to OAI, and the questionnaire did not provide any specific information about OAI. Indeed, our goal was to collect the genuine opinion of each participant about their own vision of OAI. Moreover, our dissemination channels ensured that our survey would reach subjects who (despite being highly educated) hardly possess expertise in OAI-related themes. We collected a total of 549 valid responses during Sep.–Dec. 2023.

Analysis. After collecting our responses, we split the analysis into a quantitative and a qualitative part—the latter reliant on the dual reviewing with adjudication system. The *quantitative analysis* includes demographics, expertise, and an initial overview of participants' ideas about OAI. We additionally perform a correlation analysis to understand the relationship between technical expertise and concerns about OAI. In the *qualitative analysis*, we adhere to the constructivist grounded theory methodology [57], incorporating four iterative rounds of coding. First, an "initial coding" is carried out to identify emergent themes and concepts. Then, a more "focused coding" allows us to derive broader categories of findings. Next, an "axial coding" step delineates the relationships between categories and subcategories. Finally, the "theoretical coding" integrates these categories into a comprehensive framework.

D. Systematization of Expert Opinion (Methodology)

One of the main objectives of our SoK is to identify open research problems on OAI (ℂ3). Inspired by [58–60], we do so by leveraging the collaboration with experts from the field.

Selection of Experts. To ensure the coverage of interdisciplinary perspectives of OAI, we seek to collect the opinion of experts having diverse backgrounds in terms of: area of expertise (i.e., AI, security, privacy, information systems), application domain (research or practice), institution (academia or industry), as well as gender and work experience (junior or senior). In our "recruitment" process involving private communication (with no priming), we have converged on a set of 12 experts. Eight experts stem from academia, and four from industry; eight are males, and four are females; five have a background in both security and AI, three in security, two in information systems, one in AI, and one in privacy. We did not know any expert's thoughts about OAI beforehand.

Opinion Collection and Analysis. We could have increased participation by, e.g., using interviews—but this would have introduced bias in the data collection phase [61]. Hence, to collect the opinion of experts in a bias-free way, we proceed in three steps. First, we ask experts to fill in the questionnaire for the non-expert survey (§II-C), which we have slightly modified by adding one open question requesting input about "open problems on OAI." After all 12 experts completed the questionnaire, we send them a draft of the paper and ask them to *(i)* read the paper, and *(ii)* write a paragraph of 300–500 words to describe three "open problems of OAI" based on their impressions from the draft paper as well as perception of the field. Each expert had 3 weeks to perform this task. Finally, having collected all expert responses (reported verbatim in Appendix D), we systematically analyze them quantitatively and qualitatively. Our goal is twofold: to infer open problems in OAI from the experts' input, and to compare the experts' initial input (provided *before* reading the draft paper)[3] with their final opinion (formulated *after* reading the draft paper).

Rationale. The reason why we ask our experts to provide their opinion both before and after having read our paper is twofold. On the one hand, we want to see if our systematization (and corresponding findings) "poisoned their mind": if so, this means that one of our objectives (i.e., raising awareness on some concealed aspects of OAI) was met. On the other hand, we want to analyze their opinion after providing them with *(i)* a comprehensive overview of OAI which *(ii)* aligns with our notion of OAI: indeed, the notion of OAI has been used in various way by prior work (see Fig.16), and some experts may see it differently. Hence, asking the experts to read our SoK ensures that their opinion has maximum usefulness for the sake of defining open problems. Nevertheless, we showed our manuscript to the experts after integrating and systematizing their statements in our paper, asking for further feedback (which has been used to substantially enhance this SoK).

E. OAI Assessment Checklist (Original Contribution)

We now present the method used to derive the checklist underpinning our contribution ℂ2.

> **Motivation:** Given the vast space of AI use cases it is *impossible* to identify all works that may—explicitly or implicitly—describe a *potentially offensive* AI application. Hence, our aim is to *provide the means* for systematization of knowledge in the field of OAI based on the "snapshot" of the current OAI landscape obtained as a result of our literature search (§II-A) and analysis of InfoSec events (§II-B). To this end, we develop a consistent set of criteria, implemented as a checklist, that can be matched against any relevant work (past or future). To ensure the sustainability of our checklist, we have developed a tool [21] for downstream research.

Our checklist revolves around three fundamental questions. Two are inspired by prior work [20, 62], and serve to align our SoK with existing summaries; whereas one is driven by the overarching goal of our paper. Let us describe them.

[3] We analyzed the experts' responses to the survey in the same way as we did for the non-expert survey (§II-C) and we also use our checklist (§II-E).

*1) **"What is the OAI use-case?"**:* This question, inspired by [20], enables the *systematization of the OAI use cases*. To address this question, we follow three steps. **(I)** We identify the (primary) OAI use-case, i.e., the "context" in which AI is used offensively. We begin by scrutinizing whether the work can be mapped to any of the scenarios covered in MITRE [63], and specifically in ATT&CK for Enterprise [64]. If we cannot find any match, we consider other MITRE matrices: Mobile [65] and ICS [66]. If no valid match can still be found, we assign a special category. This is common for works envisioning, e.g., attacks against privacy or society, cyberwar, or considering autonomous agents—all of which being scenarios not covered by MITRE. **(II)** We infer the overarching *purpose* of the OAI-related work. We consider the following three categories: offensive security (🛡), denoting cases wherein AI is used in, e.g., penetration testing; hacking assistant (⋆), denoting cases wherein AI is used to facilitate some procedures that could be maliciously abused; and novel attack (†), denoting works that propose full-fledged attacks reliant on AI. **(III)** We examine if the work addresses mitigation of the potential harm caused by OAI. Specifically, for works showcasing novel attacks (†), we assess whether there is an evaluation of potential defensive mechanisms; whereas for works using AI for offensive security (🛡), we inspect whether there is any statement warning downstream users that the proposed method could be maliciously exploited; we do both of these checks for works wherein AI is used as hacking assistant (⋆).

*2) **"What is the target of OAI?"**:* This is an original question of our SoK, and serves to explore the impact of OAI on the three stakeholders indicated in Fig. 1. We address this question in three steps. **(I)** We discriminate whether the OAI use-case *targets* a human (e.g., a privacy violation [67]), a system (e.g., CAPTCHA cracking [68]), or both (e.g., a phishing scenario wherein both the detector and end-users must be deceived [69]). **(II)** For works considering attacks against *systems*, we determine whether the evaluation entailed a "real" system (e.g., an operational product) or a "toy" system (e.g., a simplified version of a real system [70]). **(III)** To understand the potential relevance of the social perspective within each paper, we count the occurrences (which is a well-known practice to objectively study the focus of a document [71]) of the terms "society," "social," "societal," and "socio."

*3) **"What is the cost/benefit of OAI?"**:* This question stems from the consideration that real attackers are primarily driven by economical motives. We address this question through three steps, mostly inspired by [62]. **(I)** First, we analyze if the respective work assesses potential *benefits* for attackers from using the specific AI technique. Four outcomes are considered: *(a)* explicit mentioning of "financial" benefits (e.g., monetary gains, resources saved); *(b)* quantitative analysis via ML metrics (e.g., accuracy); *(c)* a qualitative discussion; *(d)* no mentioning at all. **(II)** Next, we assess whether the *cost* of employing OAI is taken into account, considerng the same four possible outcomes as for "benefits." **(III)** Finally, we consider if the work performed a comparison with a non-AI baseline—to determine the "contribution" of using AI for offensive purposes. Three cases are possible: *(a)* quantitative comparison; *(b)* qualitative considerations; *(c)* none.

Remark: In Appendix A, we provide a low-level description of the many elements we considered when analysing each work. Such a description is to ensure scientific transparency and reproducibility, thereby contributing to the "long-term" aspect of our checklist.

III. Overview of Academic Literature on OAI

We present the first part of our primary contribution (ℂ1) by focusing the attention on academic publications. Our literature search yielded 95 papers on OAI (see §II-A).[4] In what follows, we first discuss the 79 technical papers (§III-A) and then the 16 non-technical papers (§III-B), which we analyze under the lens of our checklist (refer to §II-E).

A. Technical Papers

We present in Table I (in the Appendix) the systematization of the 79 technical papers, in accordance with our checklist (§III-A1 to §III-A3). Furthemore, in §III-A4 we analyze the technical requirements pertaining to this class of works.

*1) **OAI Use Case**:* We begin by considering the OAI use case envisioned by each work, aligning it to MITRE ATT&CK.

- *We mapped 48 papers* (61%) *to the use-cases covered by MITRE* [63]. Among these, 2 papers were mapped to the ICS matrix (one paper focusing on Evasion and another on Process Control), whereas 1 paper was mapped to the Mobile matrix (addressing Credential Access). The remaining 45 papers aligned with the Enterprise matrix. In general, among these 48 papers, most works focus on Initial Access (22%). Other common goals of OAI are Defense Evasion (9%), Credential Access (9%), and Discovery (6%). Only 4% of the papers focus on exploiting OAI for Reconnaissance: this is likely due to the fact that this step can be carried out via various well-known means (e.g., port scanning, or OSINT) which do not require OAI and which are not easily detected [141]. We could not find any paper that specifically proposed OAI techniques for Impact or Lateral Movement (some, however, do use autonomous attack agents to carry out also these operations; e.g., [37]).
- *The remaining* (39%) *papers envisioned use-cases not covered by MITRE ATT&CK.* These papers mainly address attacks on society, privacy, or focus on autonomous attack agents (which involves automating various MITRE tactics, as done in [35]). Attacks on society cover, e.g., polarizing summaries [115] or crowdturfing attacks in online review systems [131], while privacy attacks include attribute inference attacks [89], or profile matching across multiple social networks [109, 130]. We report in Fig. 4 the groups (and corresponding relationships) of the OAI use cases not covered by MITRE.

With respect to the purpose, out of 79 papers, 43 (54%) propose novel attacks (†); 30 (38%) focus on offensive security (🛡), and 6 (8%) use AI as a hacking assistant (⋆).

[4] Among the works reviewed in this SoK, only two [49, 131] are from the big four conferences (three are from collocated workshops [135, 139, 140]).

TABLE I: **Literature Review – Technical OAI Papers.** We report 79 technical OAI papers, including discipline (see Appendix E-A; CS=Computer Science; Eco=Economics; Eng=Engineering) and the number of citations as of Jan. 2024 (taken from Google Scholar), scrutinized based on our OAI checklist (§II-E). Our search (done until Nov. 2023) identified some arXiv preprints: we report the reference to their published version (even if it appeared after Nov. 2023—e.g. [72] originally uploaded on arxiv in 2021 and published in 2024). For the specific OAI use case, we map most papers to the MITRE ATT&CK Enterprise matrix and five papers to the ICS and Mobile matrices (indicated with *); the use cases not covered by MITRE are highlighted in blue. We assign each use case ("Purpose") either to the category "defense" (🛡), "assisted-hacking" (⋆), or "attack" (†). "Def.?" denotes whether the paper considered countermeasures, "Pot. Abuse" stands for "considerations of potential malicious abuse of a defensive tool". For "Targ." (target), we use 👤 to denote "humans" and ☰ for "system"; if the target is a system, we use 🤖 to denote a "toy" system, and ⚙ for a "real" system. For the cost/benefit column, 💰 denotes a monetary assessments of the costs, 🖩 a quantitative assessment, 💬 a qualitative discussion, 🚫 is no mention. For the "Code" column, an "x" denotes if the code is available, and ✎ only denotes prompts for AI models; the icon also embeds an hyperlink to the repository (if available).

				OAI Use Case				Target/Impact			Cost/Benefit			
Paper	**Year**	**Discipline**	**Cit.**	**Specific OAI Use Case**	**Purpose**	**Def.?**	**Pot. Abuse?**	**Targ.**	**Real/Toy**	**Social Persp.**	**Benef.**	**Cost**	**Base.**	**Code**
Antonelli [72]	2024	CS	5	Init.Acc.	🛡			☰	🤖	0	🖩	💬	🖩	
AlMajali [73]	2023	CS	0	Autonomous Agents	🛡		✓	☰	🤖	0	🚫	🚫	🚫	
Chen [74]	2023	CS	3	Autonomous Agents	🛡			☰	🤖	0	🖩	🖩	💬	x
Chowdhary [75]	2023	CS	0	Init.Acc.	🛡			☰	⚙	0	🚫	🚫	🚫	
Gallus [76]	2023	CS	0	Init.Acc.	⋆			☰	🤖	1	🚫	🚫	🚫	✎
Ghanem [77]	2023	CS	13	Autonomous Agents	🛡			☰	🤖	0	💬	💬	🚫	
Happe [78]	2023	CS	8	Autonomous Agents	⋆		✓	☰	🤖	0	🚫	🚫	🚫	x
Iqbal [79]	2023	CS	3	Autonomous Agents	⋆			☰	🤖	4	💬	🚫	🚫	✎
Karinshak [80]	2023	CS	17	Atk on soc.	†			👤		20	💬	🚫	💬	✎
Ozturk [81]	2023	CS	2	Disc.	⋆			☰	🤖	1	🖩	🚫	🖩	✎
Pa Pa [82]	2023	CS	9	Res.Dev.	⋆	✓		☰	🤖	3	💬	🖩	🚫	✎
Zennaro [34]	2023	CS	25	Autonomous Agents	🛡		✓	☰	🤖	3	💬	🚫	🚫	x
Auricchio [83]	2022	CS	6	Init.Acc.	🛡			☰	🤖	0	🖩	💬	🖩	
Biesner [84]	2022	CS	3	Cred.Acc.	†			☰	🤖	1	🖩	🚫	🖩	x
Cody [85]	2022	CS	15	Exfil.	🛡			☰	🤖	0	🚫	🚫	🚫	
Confido [86]	2022	Eng.	2	Autonomous Agents	🛡			☰	🤖	4	🖩	🖩	💬	
Gangupantulu [87]	2022	CS	27	Autonomous Agents	🛡			☰	🤖	1	💬	🚫	🚫	
Hu [25]	2022	CS	635	Def.Ev.	†	✓		☰	🤖	0	🚫	🚫	🚫	
Jagamogan [88]	2022	CS	2	Init.Acc.	🛡			☰	🤖	0	🖩	🚫	🖩	x
Karanatsiou [89]	2022	CS	18	Priv.atk.	†			👤		59	🚫	🚫	🚫	
Lee [90]	2022	CS	16	Init.Acc.	†			☰	⚙	0	💰	🚫	🖩	x
Li [91]	2022	CS	6	Autonomous Agents(ICS)	🛡			☰	🤖	0	🚫	🚫	🚫	
Lin [26]	2022	CS	332	Def.Ev.	†	✓		☰	🤖	0	🖩	🚫	🖩	
Nhu [92]	2022	CS	3	Autonomous Agents	🛡			☰	🤖	0	🚫	🚫	🚫	
Pagnotta [93]	2022	CS	8	Cred.Acc.	†			☰	🤖	0	🖩	🚫	🚫	
Tran [94]	2022	CS	8	Autonomous Agents	🛡			☰	🤖	0	🖩	🚫	🚫	
Yao [95]	2022	CS	0	Autonomous Agents	🛡			☰	🤖	0	🚫	🚫	🚫	
Caturano [96]	2021	CS	29	Init.Acc.	⋆			☰	🤖	0	🚫	🚫	🖩	
Erdődi [33]	2021	CS	38	Init.Acc.	🛡		✓	☰	🤖	0	🚫	🚫	🚫	x
Gangupantulu [97]	2021	CS	13	Disc.	🛡			☰	🤖	1	🚫	🚫	🚫	
Khan [98]	2021	Eco	5	Init.Acc.	†			👤+☰	⚙	4	💬	💬	🚫	
Kujanpää [99]	2021	CS	8	Priv.Esc.	†	✓		☰	🤖	0	💬	🚫	🚫	
Lee [100]	2021	CS	4	Cred.Acc.	†			☰	🤖	0	🖩	🚫	🖩	
Maeda [101]	2021	CS	49	Priv.Esc.	🛡		✓	☰	🤖	0	💬	🚫	💬	
Neal [102]	2021	Eng.	12	Procc.Contr.*	🛡			☰	🤖	0	🚫	🚫	🚫	
Sharevski [103]	2021	CS	1	Atk on soc.	†			👤		15	💬	💬	🚫	
Standen [104]	2021	CS	52	Priv.Esc.	🛡			☰	🤖	0	🚫	🚫	🚫	
Toemmel [105]	2021	CS	1	Pers.	†			👤+☰	🤖	0	🚫	🚫	🚫	
Tran [106]	2021	CS	31	Autonomous Agents	🛡			☰	🤖	0	🖩	🚫	🚫	
Al-Hababi [107]	2020	CS	9	Recon.	†			☰	🤖	11	🚫	🚫	🚫	
Bhattacharya [37]	2020	CS	14	Autonomous Agents(ICS)	🛡			☰	🤖	0	💬	💬	🚫	
Chowdhary [108]	2020	CS	49	Autonomous Agents	🛡			☰	🤖	0	💬	💬	💬	x
Halimi [109]	2020	CS	6	Priv.atk.	†			👤		36	🖩	🚫	🚫	
Hu [110]	2020	CS	74	Autonomous Agents	🛡			☰	🤖	0	🚫	🚫	🚫	
Lee [111]	2020	CS	18	Cred.Acc.	†			☰	🤖	2	💬	🚫	🚫	
Lee [112]	2020	CS	4	Cred.Acc.	†			☰	🤖	2	🖩	🚫	🖩	
Liu [113]	2020	CS	42	Init.Acc.	🛡			☰	⚙	0	💰	🖩	🖩	x
Pearce [114]	2020	CS	5	Def.Ev.	†	✓		☰	🤖	0	💬	🚫	🚫	x
Sharevski [115]	2020	CS	3	Atk on soc.	†	✓		👤+☰	🤖	13	🚫	🚫	🚫	
Shu [116]	2020	CS	46	Def.Ev.	†	✓		☰	🤖	1	💬	🚫	🚫	
Song [117]	2020	CS	33	Def.Ev.	†	✓		☰	⚙	0	🖩	🚫	🚫	x
Valea [35]	2020	CS	27	Autonomous Agents	🛡			☰	🤖	0	💬	💬	🚫	
Yu [118]	2020	CS	8	Disc.	†	✓		☰	⚙	0	💬	💬	🖩	
Basu [119]	2019	CS	0	Init.Acc.	†	✓		👤+☰	⚙	8	🚫	🚫	🖩	
Cecconello [120]	2019	CS	12	Recon.	†	✓		👤+☰	⚙	5	💬	💬	🚫	
Chung [121]	2019	CS	31	Evas.*	†	✓		☰	🤖	0	💬	💬	🚫	
Das [122]	2019	CS	115	Recon.	†			☰	⚙	0	🖩	🚫	🖩	x
Ghanem [123]	2019	CS	104	Autonomous Agents	🛡			☰	🤖	0	💬	💬	🚫	
Tshimula [124]	2019	CS	10	Priv.atk.	†			👤		24	🚫	🚫	🚫	
Yu [68]	2019	CS	26	Init.Acc.	†			☰	⚙	0	🚫	💬	🚫	
Zhang [125]	2019	CS	18	Cred.Acc.*	†	✓		☰	⚙	0	🚫	🚫	🚫	
Anand [126]	2018	CS	19	Cred.Acc.	†	✓		☰	🤖	1	💬	🚫	🖩	
Bahnsen [127]	2018	CS	78	Def.Ev.	†	✓		☰	🤖	0	🖩	🚫	🖩	x
Kronjee [128]	2018	CS	51	Init.Acc.	†	✓		☰	⚙	0	🖩	💬	💬	x
Rigaki [129]	2018	CS	133	Def.Ev.	†	✓		☰	⚙	0	💬	🚫	🚫	
Zhuo [130]	2018	CS	173	Priv.atk.	†			👤		29	🖩	🚫	🖩	x
Yao [131]	2017	CS	210	Atk on soc.	†	✓		👤+☰	🤖	24	🖩	💬	🖩	
Anderson [132]	2016	CS	230	C2	†	✓		☰	🤖	0	🚫	🚫	🚫	x
Ceccato [36]	2016	CS	45	Init.Acc.	🛡			☰	⚙	0	💬	💬	🖩	
Grieco [133]	2016	CS	276	Disc.	🛡			☰	⚙	0	🚫	🚫	🖩	x
Freitas [134]	2015	CS	181	Atk on soc.	†			👤+☰	⚙	230	🖩	🚫	🚫	
Bursztein [135]	2014	CS	178	Init.Acc.	†	✓		☰	⚙	0	🚫	💬	🖩	
Adali [136]	2012	CS	151	Priv.atk.	†			👤		39	🚫	🚫	🚫	
Malhotra [137]	2012	CS	307	Priv.atk.	†			👤		53	🚫	🚫	🚫	
Sumner [138]	2012	CS	423	Priv.atk.	†			👤		55	🚫	🚫	🚫	
Goldbeck [23]	2011	CS	852	Priv.atk.	†			👤		35	🚫	🚫	🚫	
Yamaguchi [139]	2011	CS	212	Disc.	🛡		✓	☰	⚙	0	🚫	🚫	🚫	
Bursztein [140]	2009	CS	103	Init.Acc.	†	✓		☰	⚙	0	🖩	🚫	🚫	
Golle [49]	2008	CS	394	Init.Acc.	†	✓		☰	⚙	0	🚫	🚫	🚫	

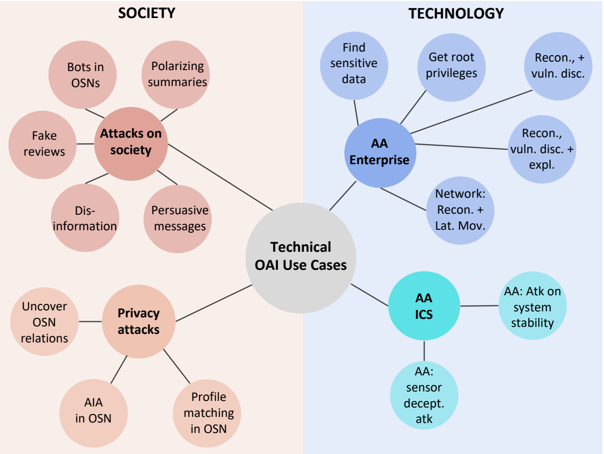


Fig. 4: **OAI use cases not covered by MITRE ATT&CK (technical papers).** Some focus on society and privacy (OSN=online social networks).

Some works focusing on novel attacks (†) consider potential defenses against the proposed attacks: e.g., [75] considers Web Application Firewalls to protect the targeted application. Only half (51%) of the novel attack (†) papers, however, propose/evaluate a potential countermeasure, which is alarming. Moreover, for those papers that focus on offensive security (🛡), we found that only 14% of these considered the possibility that their tool could be used for malicious purposes. Notably, only one work ([139]) did so before 2020. We believe that the publication of the "Autonomous Weapon Open Letter" from Future of Life [142] in 2016 may have encouraged more works to acknowledge potential abuse of AI-based tools (for instance, this letter was explicitly mentioned in both [33, 34]). Nonetheless, to shed further light on this aspect, *we reached out to the authors of 25 offensive security* (🛡) papers in which we found no explicit mention of potential abuse of the proposed AI-based tool, inquiring if such a scenario would open up potential security concerns. We received a response from the authors of 11 of these papers (44% response rate): while all responses did not deny such a possibility, the overall sentiment is that a considerable amount of effort is necessary to exploit the proposed AI-based tool for real-world attacks.

Lessons Learned: (1) Many reported OAI techniques transcend the use-cases covered by MITRE ATT&CK, which was primarily used in prior summaries, such as [20]. **(2)** Many works reporting OAI techniques do not adequately consider countermeasures. **(3)** Even ethical statements that merely acknowledge potential abuse risks for considered AI techniques are quite uncommon in prior OAI research.

*2) **Target/Impact**:* For the second set of criteria in our checklist, we first assess whether the reported OAI techniques target a human (👤), a system (🖥), or both; whether attacks against systems target "toy" (🤖) or "real" (⚙) systems. Among the analyzed 79 papers, 62 (78%) consider attacks against systems, while 10 (13%) propose attacks against humans, and 7 (9%) use AI to attack both systems and humans. For the 69 papers that attack systems, 70% consider "toy" (🤖) systems, whereas only 21 papers target "real" (⚙) systems.[5] Among these 21, the majority (52%) use OAI for "Initial Access" (more precisely, on phishing, web applications, and CAPTCHAs). Then, we assess the emphasis given by the paper on the societal aspect of its contribution, and we counted the number of occurrences of the words "society," "social," "societal," or "socio" in the main text of the paper. We observed that 49 papers (62%) *never mention any of these terms*. Among those that do use these terms, most works belong to the "privacy" domain (e.g., [134] mentions these terms 230 times). Finally, for the 17 papers considering attacks against humans, we investigated whether they carried out any validation (e.g., a user study) on real people. None of these 17 papers did. A likely explanation for this is the difficulty to carry out such experiments due to, e.g., ethical concerns.

Lessons Learned: (1) Most prior technical papers have targeted systems rather than humans; however, attacks against *real systems* are scarce. This finding (which echoes [62]) suggests that the impact of OAI techniques on real systems may still be unknown. **(2)** Most works do not emphasize the impact of their findings on our society. This may suggest that the societal impact of OAI on our society is difficult to estimate in this domain.

*3) **Cost/Benefit**:* To shed light on the practicality of the identified techniques, we examined whether the authors considered attackers' economic incentives for applying the proposed method (benefits) or highlighted potential costs that may deter attackers from doing so. Among our 79 papers, 32 (41%) did not provide any analysis (🚫) of the attacker's benefits when leveraging AI: a typical conclusion in these papers is that the "proposed method works." Among the remaining 47 papers, 23 (29%) provided a qualitative evaluation (💬) of the benefits, and 22 (28%) even a quantitative evaluation (🖩). Only 2% explicitly mentioned (💰) monetary benefits or time saved according to metrics that go beyond sheer accuracy and precision, e.g. [113]. Concerning the *costs*, 58 (73%) papers do not provide any analysis, while 17 (22%) only provide a brief qualitative evaluation. The remaining 4 papers (5%) provide a quantitative analysis. In general, no paper (among our 79 technical papers) has precisely quantified the required investment to launch the attack, or the return on such investment if the attack were successful. Finally, another measure of the attack practicality is whether the same objective could be achieved without AI. The majority of papers (65%) did not consider any non-AI-baseline for comparison.

Lessons Learned: The economical aspect tends to be neglected by most technical papers. Such a finding raises a question, to what extent OAI represents a tangible threat in practice and if so, what threat actors (from individuals to state-sponsored groups) are likely to deploy such techniques. Future work should take such cost factors into account.

[5] For example, the authors of [120] attack a Voice-over-IP software (Skype) which we consider as a "real" system; whereas the authors of [132] attack a DGA detector envisioned by prior work ([143, 144]), i.e. a "toy" system.

TABLE II: **Literature Review – Non-Technical OAI Papers.** We report 16 non-technical OAI papers (scrutinized under the same criteria as those in Table I).

				OAI Use Case				Target/Impact			Cost/Benefit			
Paper	**Year**	**Discipline**	**Cit.**	**Specific OAI Use Case**	**Purpose**	**Def.**	**Pot. Abuse?**	**Targ.**	**Real/Toy**	**Social Persp.**	**Benef.**	**Cost**	**Base.**	**Code**
Dall'Agnol [145]	2023	Soc. Sciences	12	Atk.in (cyber) war	†	✓		👤+≣		0	⊘	⊘	⊘	
De Angelis [146]	2023	Medicine	166	Atk.on society	†	✓		👤		7	💬	💬	⊘	
Illiashenko [147]	2023	CS	4	Atk.on society	†	✓		👤		2	💬	⊘	⊘	
Pashentsev [148]	2023	Soc. Sciences	0	Atk.on society	†	✓		👤+≣		161	💬	⊘	⊘	
Rickli [149]	2023	Soc. Sciences	3	Atk.in (cyber) war	†			≣		0	⊘	⊘	⊘	
Hao [150]	2022	CS	0	Autonomous Agents	🛡		✓	👤+≣		1	⊘	⊘	⊘	
Kasim [151]	2022	CS	0	Autonomous Agents	†	✓		👤+≣		0	🖩	🖩	🖩	
McIlroy-Young [152]	2022	CS	7	Atk.on society	†	✓		👤		17	💬	💬	⊘	
Nica [153]	2020	Soc. Sciences	1	Atk.in (cyber) war	†			👤+≣		18	💬	💬	⊘	
Skeba [154]	2020	CS	11	Priv.atk	†	✓		👤		28	⊘	⊘	⊘	
Easttom [155]	2019	CS	5	Atk.in (cyber) war	†			👤+≣		0	⊘	⊘	⊘	
Burton [156]	2019	CS	5	Atk.in (cyber) war	†			≣		0	💬	⊘	⊘	
Burton [157]	2019	CS	47	Atk.in (cyber) war	†			👤+≣		15	💬	💬	⊘	
Giaretta [158]	2019	CS	11	Init.Acc.	†			👤		4	⊘	⊘	⊘	
Maus [159]	2015	CS	7	Atk.on society	†	✓		≣.		64	💬	⊘	⊘	
Guarino [160]	2013	CS	22	Atk.in (cyber) war	†			👤+≣		0	💬	💬	⊘	

4) ***Technical Requirements:*** We conclude this section by analyzing the technical requirements of the 79 "technical" papers. Specifically, we address the question "What is the degree of technical effort required to set up the proposed OAI tool?" At a high level, we proceed as follows.

- First, we examine whether the tool is based on a pre-existing ML model (e.g., ChatGPT) or if the ML model must be developed from scratch. Only 9 (11%) papers rely on a pre-existing model (e.g., [98] used a pre-trained Large Language Model to generate spear-phishing emails with the intent of deceiving the system and the user).
- If the ML model must be developed from scratch (which is the case for 70 papers out of 79) we scrutinize the availability of the data required to train such an ML model. 25 papers (36%) rely on publicly available data (e.g., benchmarks, such as [84]); for 6 papers (9%), the authors needed special access rights to obtain their training data (e.g., [121]); the remaining 39 papers (57%), entailed creation of a custom training dataset (e.g., [121]).
- Finally, we review the reproducibility of the implementation. Only 17 (22%) papers released their source code, and 5 papers (6%) release the exact prompts used to realize the attack. In contrast, 57 papers (72%) do not provide such low-level details (a result which echoes [62, 161]).

A detailed explanation of this analysis is provided in Appendix A-D (and these results are also shown in Table IV).

> **Lessons Learned:** Most technical papers implement their OAI tool from scratch (and few release their code/data), suggesting that implementing/launching the attack by third parties is not trivial in practice. However, we can expect this trend to change given the increasing availability of LLM—which could benefit both researchers and attackers.

B. Non-technical Papers

We now analyze the 16 "non-technical" papers (in Table II in the Appendix) through our checklist and discuss our findings.

1) ***OAI Use Case:*** We could only map one paper [158] to MITRE. The remaining 15 (94%) papers consider use cases not covered by MITRE. For instance, [151] uses game theory to analyze whether "human defenders" can withstand "AI attackers," and conclude that well-trained AI agents are almost impossible to beat. Intriguingly, 7 papers (46%) focus on *cyber warfare* (not covered by MITRE). For instance, [155] provides theoretical arguments on how AI could be used to develop malware bypassing the detection mechanisms of the attacked entity in a cyber war. In terms of purpose, 15 papers envision using AI for novel attacks (†), and 1 for offensive security (🛡).

2) ***Target/Impact:*** Most papers (13, 81%) consider OAI targeting humans (👤), and the overall occurrence of society-related terms is higher than for technical papers (38% of papers in Table II have ten or more occurrences, compared to 18% for those in Table I). Finally, none of these 16 papers carry out user studies with real humans to validate any given hypothesis.

3) ***Cost/Benefit:*** Most papers provide a shallow analysis of benefits (38% ⊘, 56% 💬) and costs (63% ⊘, 31% 💬).

> **Lessons Learned:** Non-technical papers on OAI put a greater emphasis on the human perspective and envision scenarios not covered by MITRE, such as cyber warfare. However, these works do not carry out user studies to validate their hypotheses or assess the opinion of real people. Lack of such a validation may either over- or under-estimate the relevance of the envisioned OAI scenario to our society.

IV. OAI in InfoSec Briefings

We identified 38 non-academic works (also known as "briefings") related to OAI at BlackHat and DefCon. We first analyze these 38 briefings (§IV-A), and then compare them with the 95 papers from academic literature (§IV-B).

A. Analysis

To finalize our primary contribution (ℂ1), we assess our 38 briefings through our checklist (§II-E), and show the results in Table III (in the Appendix), which also reports briefings related to a scientific paper (which occurs for 9 out of 38 briefings—two of which [120, 127] are also included in Table II). The first briefing on OAI we found (discussing how to "hack human desire") dates back to 2008 [50]; many briefings on OAI appeared in 2023—likely due to the rollout of ChatGPT.

1) ***OAI Use Case:*** Most briefings (25, 69%) can be mapped to MITRE, for which "Initial Access" (11, such as using LLM to write phishing emails [164]) and "Reconnaissance" (8, e.g., via side-channel [167, 180, 187]) are the most prominent use cases. The remaining 13 briefings are not covered by MITRE. These briefings focus mostly on attacks against society (9, such as using AI for virtual kidnapping [168]) and privacy (3, such as deanonymizing developers based on their code [184]).

TABLE III: **Analysis of InfoSec Briefings.** We report 38 industrial "InfoSec" briefings from BlackHat and DefCon (scrutinized under the same criteria as the works in Table I). Column "Acad.?" denotes whether a briefing also has a corresponding publication (we denote * with preprints).

		OAI Use Case				Target/Impact			Cost/Benefit				
Author	Year	Specific OAI Use Case	Purpose	Def.?	Pot. Abuse?	Targ.	Real/Toy	Social Persp.	Benef.	Cost	Base.	Code	Acad.?
Scheiner [162]	2023	Atk. on society	†	✓		👤		7	💬	💬	💬		
Canham [163]	2023	Atk. on society	†	✓		👤		11	💬	💬	💬		
Heiding [164]	2023	Init.Acc.	†	✓		👤+🗄	⚙	0	🖩	💬	🖩		[165]*
Herbert-Voss [166]	2023	Init.Acc.	⋆			🗄	⚙	2	💬	💬	💬		
Waligóra [167]	2023	Recon.	†			🗄	⚙	0	💬	🪙	🚫		
Gibson [168]	2023	Atk. on society	†	✓		👤		4	💬	💬	🚫		
Zror [169]	2023	Recon.	†			👤		7	💬	💬	💬		
Xing [170]	2022	Init.Acc.	†	✓		👤+🗄	🤖	2	🖩	🚫	🖩	x	
Chi [171]	2022	Init.Acc.	🛡		✓	🗄	⚙	0	💬	🚫	💬		
Lim [172]	2021	Init.Acc.	†	✓		👤+🗄	⚙	8	🖩	💬	🖩		
Lohn [173]	2021	Atk. on society	†	✓		👤		3	🖩	🖩	🚫		
Tully [174]	2020	Atk. on society	†	✓		👤+🗄	⚙	20	💬	💬	🚫		
Basu [175]	2020	Init.Acc.	†			👤		2	🚫	🚫	🚫	x	
Sharma [176]	2020	Disc.	🛡			🗄	⚙	2	💬	🚫	🚫		[177]
Takaesu [178]	2019	Autonomous Agents	🛡		✓	🗄	⚙	0	🚫	🚫	🚫	x	
Botwicz [179]	2019	Recon.	🛡		✓	🗄	⚙	0	🚫	🚫	🚫	x	
Bursztein [180]	2019	Recon.	†			🗄	⚙	0	💬	💬	🚫	x	
Ding [181]	2019	Init.Acc.	🛡			🗄	⚙	0	💬	🚫	🚫		
Price [182]	2019	Atk. on society	†	✓		👤+🗄	⚙	1	🚫	🚫	🚫	x	
Bahnsen [183]	2018	Dev.Ev.	†	✓		🗄	🤖	1	🖩	🚫	🖩	x	[127]
Greenstadt [184]	2018	Priv.atk	†			👤		0	🖩	🚫	🚫		[185]
Kirat [186]	2018	Def.Ev.	†	✓		👤+🗄	⚙	2	🖩	🚫	🚫		
Perin [187]	2018	Recon.	†			🗄	⚙	0	🖩	🚫	💬		
Gomez [188]	2018	Priv.atk	†			👤		4	🚫	🚫	🚫	x	
Anderson [189]	2017	Def.Ev.	†	✓		🗄	🤖	0	🖩	🖩	🚫	x	[190]
Lain [191]	2017	Recon.	†	✓		👤+🗄	⚙	0	💬	💬	🚫	x	[120]
Morris [192]	2017	Init.Acc.	†			🗄	⚙	22	🚫	🚫	🚫		
Tully [193]	2017	Atk. on society	†	✓		👤+🗄	⚙	69	💬	💬	🚫		
Singh [194]	2017	Recon.	†	✓		👤+🗄	⚙	13	🚫	🚫	🚫		
Polakis [195]	2016	Init.Acc.	†	✓		🗄	⚙	0	🪙	🪙	🖩		[196]
Argyros [197]	2016	Init.Acc.	🛡			🗄	⚙	0	🚫	🚫	🚫	x	[198]
Seymour [199]	2016	Init.Acc.	†	✓		👤+🗄	⚙	22	🖩	🖩	🖩		
Wolff [200]	2016	Exfil.	†			🗄	⚙	0	🚫	🚫	🚫		
Bursztein [201]	2014	Atk. on society	†			👤+🗄	⚙	0	💬	🚫	🚫	x	
Fu [202]	2014	Priv.atk	†	✓		👤+🗄	⚙	0	💬	🚫	🚫		[203]
Vanned [204]	2013	Res.Dev.	†			🗄	⚙	0	🪙	🖩	🖩	x	
Espinhara [205]	2013	Recon.	†			👤+🗄	⚙	41	🚫	🚫	🚫	x	
Clarke [206]	2008	Atk. on society	†			👤		0	🖩	🚫	🚫		

Overall, only one briefing [166] used AI as a hacking assistant (⋆). In contrast, 6 briefings envisioned an offensive security application (🛡), half of which explicitly mention that the proposed tool can be maliciously exploited also by attackers. The remaining 31 briefings proposed a new attack reliant on AI (†): among these, 18 consider a countermeasure.

2) ***Target/Impact:*** 9 briefings consider OAI applications targeting humans (👤), 16 a system (🗄), and 13 both. Among those that target systems, 26 attack real systems, and 3 toy systems (e.g., [170] generates AI-synthesized speech samples and tests them against three fake-voice detectors proposed by prior academic literature—i.e., a "toy" system). The term "society" is never mentioned in 18 (47%) briefings; the most occurrences are found in [193] which considers using AI to hide data in images posted on social networks. Only one briefing entails a user study: [164] compares the performance of humans and LLM to write phishing emails, and tests which "author" was more effective at fooling end-users.

3) ***Cost/Benefit:*** Only 6 (16%) briefings make a clear analysis (🖩 or 🪙) of the *costs* required to realize the attack: for instance, [167] tests the attack on real hardware, quantifying the costs of the proposed attack as: "a laptop + $100 PicoScope (software)." In contrast, 21 briefings (55%) do not mention the cost/benefit aspect at all, and 11 (29%) make a shallow qualitative assessment. The opposite holds for the *benefits*: only 10 briefings (26%) do not mention this aspect, whereas 13 (34%) perform actual measurements (🖩 or 🪙) and 15 (40%) make some qualitative analyses.

B. Comparison: Scientific Literature vs InfoSec Briefings

Intriguingly, our search revealed that the initial efforts on OAI in both literature and InfoSec date back to 2008. While academic literature has consistently tackled OAI since then, interest in InfoSec venues only started to increase in 2013.

Let us elucidate the main similarities and differences among these two sources. In regard to "*1) OAI Use Case*," technical papers from academia are more similar to Infosec briefings than to non-technical works. Indeed, 69% of InfoSec briefings and 61% of technical papers can be mapped to MITRE (albeit the specific use-cases differ), while this holds for only 6% of non-technical papers. At the same time, InfoSec briefings and technical papers have not focused on cyber warfare, while 46% of non-technical papers examined this topic. One notable difference, however, is that only 6% of technical papers focus on attacks on society—whereas the share is 24% and 31% for InfoSec briefings and non-technical papers, respectively.

Related to "*2) Target/Impact*," 78% of technical papers consider attacks against systems, compared to 42% for InfoSec briefings, which consider humans in the remaining attacks (either isolated or combined with systems). However, InfoSec briefings consider more attacks on real systems (90%) compared to technical papers (30%). For *3) Cost/Benefit* the economic aspect is mostly neglected by academic works, proposing attacks that need to be built from scratch and with custom datasets (57%). Among InfoSec briefings 16% make a clear analysis of the costs required to carry out the attack, and 34% make actual measurements of the benefits. While academic technical works encompass compelling use cases, InfoSec briefings tend to provide a more comprehensive examination of risks (*in the wild*). Although this observation appears intuitive, integrating both perspectives drives a comprehensive assessment of the risks posed by OAI to society.

Lessons Learned: Most OAI use cases of InfoSec briefings are not discussed in research papers. Only 47% briefings consider countermeasures or emphasize that attackers can use their tool. These results highlight blind spots exploitable by attackers, to be addressed by future research.

V. User Survey: what do laypeople think of OAI?

We now present the results of our user survey with non-experts, revealing what laypeople think about the offensive potential of AI. We first describe the demographics of our participants (§V-A), and then present the results (§V-B)

A. Demographics

We received 570 responses but removed 21 because they were from underage participants or clearly not informative. Hence, our results reflect the opinion of 549 individuals.

Overall, 63% of our respondents identify themselves as "male," 36% as "female" and less than 1% as "other." Our respondents are from 42 countries, mostly from Europe (70%), North America (21%), and Asia (7%). Over 90% are between 18–54 years old, and 2% are over 65. With respect to educational qualifications, 80% of participants hold at least a Bachelor's degree. Employment status among the respondents varies, with 83% employed (full or part-time), 15% students, and 2% retired or unemployed. 75% of the participants are engaged in IT-related work. In terms of knowledge of cybersecurity (or AI), 23% (17%) consider themselves as beginners, 43% (53%) as intermediate, and 34% (30%) as advanced or experts. Additional demographic details are provided in Appendix B.

Compared to the OECD population, our sample has fewer women and individuals over 55. However, we appreciate that our sample consists of highly educated individuals, of which only a minority consider themselves as experts in AI or security. Therefore, even though we cannot claim representativeness of the world's population, our sample is likely to provide valuable insights for the goal of our SoK. To our knowledge, this is the first survey of this kind on OAI, hence its findings are useful (also) for future studies.

B. Results

We systematically analyze our results quantitatively (§V-B2) and qualitatively (§V-B1) before drawing our conclusions.

1) ***Quantitative Analysis****:* We report the results of our binary questions in Figure 5. Let us analyze it at a high-level.

- *"Have you thought about OAI?"* The majority (525, 96%) have already considered that AI could be used for malicious purposes. However, the remaining 24 (4%) have never considered AI's offensive potential. Among these, 14 people work in an IT-related field, and 13 of these have at least intermediate knowledge in cybersecurity or AI.
- *"Are you concerned about OAI?"* Among the 525 participants that have considered the offensive potential of AI, 442 are concerned about it (and 83 are not concerned).
- *"Do you think that AI will harm you?"* A slight majority (52%) believe that AI may personally harm them. Intriguingly, nearly 40% of the 442 participants who are concerned about OAI do not believe that AI will harm them; whereas 18% of the 83 participants that are not concerned about OAI believe that AI will harm them.

We have carried out correlation analyses (in Appendix B) to determine whether an individual's background affects their views on OAI. We found no correlation between having a job in IT and having concerns about OAI or believing that AI may inflict personal harm. We found a weak positive correlation between the level of expertise and concerns about OAI: in particular, knowledge of AI has a stronger impact than knowledge in cybersecurity (Pearson's ρ=0.27 vs. 0.11). Yet, we found no correlation between the expertise in either cybersecurity or AI and the belief that AI will inflict harm.

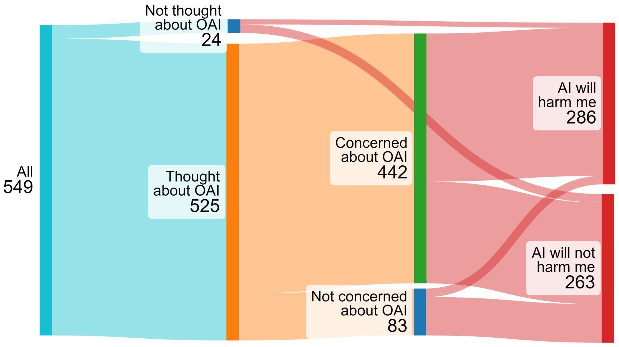


Fig. 5: **Quantitative results (laypeople).** Sankey chart of the closed questions.

2) ***Qualitative Analysis****:* We review the answers to the open questions wherein we ask participants to provide reasons why they are (or are not) concerned about OAI. Here, we present the results of our coding-based analyses (discussed in Appendix C). Every respondent could mention more than one source of concern (or lack thereof).

- *"What are you most concerned about OAI?" (442 respondents)* Most respondents (141) are concerned about the broad category of "cyberattacks" or provide unspecific generic concerns (24). Among those that do provide clear scenarios, 98 mention "spread of misinformation," 61 "deepfakes," 21 about "military applications of AI" and 19 about "privacy attacks." Some even mentioned concerns that are orthogonal to our vision of OAI: for instance, many are concerned about improper usage of AI by its developers (i.e., 20 mentioned "losing control of AI," 12 "data misuse," 10 "unintended errors," 9 "regulatory deficiencies," 8 "negligent AI development") including "ethical concerns" (19); whereas others are concerned about "job displacement," (16) "reduction in human learning" (15), or "AI surpassing human performance" (14).
- *"Why are you not concerned about OAI?" (83 respondents)* Most respondents (26) simply do not express any clear reason for their lack of concerns. Many (21) believe that current protection mechanisms are enough to handle the threat of OAI. Some (9) are more concerned about human-centered issues (e.g., ethics) of AI. A minority believes that AI is not yet mature (5) or that the benefits of AI

outweigh potential issues (8). Some even suggested proactivity, stating that "Just being concerned is not helpful" or "we should be prepared and not just concerned." These results echo those of a 2023 survey among 199 participants, wherein 35% of participants stated that generative AI will not provide an advantage to attackers or defenders [207].

We further scrutinize the responses of those (141) participants that mentioned concerns about "cyberattacks." While 58 do not provide additional details, 78 mention use-cases that we were able to map to MITRE ATT&CK: specifically, 26 consider "reconnaissance," 20 "resource development," 18 "initial access," 9 "defense evasion," 2 "credential access" and 1 "lateral movement." Finally, 5 are concerned about autonomous attacks or adversarial ML attacks (which are unrelated to OAI).

> **Lessons Learned:** Over 80% of our respondents are concerned about OAI. However, our sample may have misconceptions about "offensive AI," since some concerns relate to problems that are orthogonal to how evildoers may use AI to cause harm. This suggests that non-experts may be oblivious of the offensive potential of AI, thereby underscoring the necessity of proper awareness campaigns.

VI. Expert Opinion: What is the future of OAI?

In this section, we report the systematization of the expert opinions (refer to §II-D). We begin by presenting their responses to the survey (§VI-A); then, we summarize their input after they reviewed this paper (§VI-B); finally, we compare their opinions before and after having read our paper (§VI-C).

A. Expert Survey (opinions before reviewing this paper)

Recall that we first inquired the 12 experts to participate in a (slightly modified) version of our survey. Hence, their responses reveal their *unbiased opinion* on OAI. In terms of demographics, 5 (42%) identify as an expert in AI, 4 (33%) as advanced, and 3 (25%) as intermediate; whereas 9 (75%) deem to be an expert in cybersecurity, and 3 (25%) identify as having advanced knowledge in cybersecurity. This is in line with our expected target of "experts" in these fields.

*1) **Quantitative Analysis**:* We first report the results to the binary questions (as we did in §V-B1). All 12 experts have already considered that AI could be used for malicious purposes, and also all experts are concerned about the offensive potential of AI. For the last question, 9 (75%) experts think that AI could harm them personally, while 3 (25%) think otherwise.

*2) **Qualitative Analysis**:* Next, we qualitatively analyze the answers to the open questions—starting from the one that was also included in the survey with the general population.

- *"What are you most concerned about OAI?"* The most prevalent concerns are "cyberattacks" (4), and the "speed, automation and ease of use" (4). Some experts are concerned about "privacy attacks" (2), "deepfakes" (1), and the "spread of misinformation" (1).
- *"Can you think of (or do you know of) some ways in which AI can be used offensively?" [expert-only]* We mapped the answers to the OAI use cases of our checklist (§II-E). Most experts (6) highlighted multiple use cases. For MITRE ATT&CK: 7 experts mentioned "initial access," 3 "resource development," and 1 "reconnaissance." 4 experts highlighted attacks against society, and 2 experts autonomous attacks. None of the experts mentioned privacy attacks.
- *"What are some means that can be used to counter AI-powered cyberattacks?" [expert-only]* Most experts (8) stated multiple countermeasures. Intriguingly, 4 experts recommend to use AI-powered countermeasures, whereas 6 think that AI is not necessarily required in such defenses. Additional solutions include: user education and awareness (4); mechanisms to recognize "AI behavior" (4), such as CAPTCHAs; data anonymization (1); as well as banning generative AI for public figures (1). Finally, 2 experts think that OAI threats can be tackled through the same mechanisms as traditional cyber attacks.
- *"Which stakeholders should be responsible for implementing/realizing/advertising such countermeasures?" [expert-only]* To analyze these answers, we grouped the mentioned stakeholders in: system providers (industry, technical service providers, vendors); sovereign entities (government, institutions, regulations); and generic humans (individuals, users). Among our experts, 6 consider system providers as the primary responsible parties; whereas 5 indicate system providers *and* sovereign entities; one considers everyone as equally responsible (i.e., system providers, sovereign entities, and the general human population).

We will draw some comparisons between these results and those of the survey among the general population in §VII-A.

B. Expert Statements (after reading our paper)

We now focus on the statements that the 12 experts contributed after reviewing our draft paper. We first objectively analyze these texts via natural language processing (NLP) techniques (§VI-B1). Then, we coalesce the experts' opinions into 7 open problems and 3 fundamental concerns about OAI (§VI-B2), serving as a basis for future work on OAI (ℂ3).

*1) **Preliminary Analysis**:* Overall, the statements written by our 12 experts span across ≈5k words and ≈35k characters. To provide an objective foundation for a systematic assessment, we carry out a preliminary analysis for which we rely on well-known NLP techniques for text mining. Specifically, we first extract the 20 most common bi-/tri-grams across the entire statements; then, we perform a more fine-grained analysis and apply KeyBERT [208] to extract the 5 most relevant keywords for each expert statement. Finally, we apply topic modeling via BERTopic [209] to identify the most relevant topics across the entire statements. We provide in the Appendix D-B a more low-level description of these procedures (including how they work and why they are relevant), as well as the detailed results (the full source code is available in our repository [54]). At a high-level, our analyses with BERTopic revealed that certain topics were more prevalent than others. For instance, words such as "bias," "cognitive," and "exploit," had a lot of weight, suggesting the topic of "cognitive bias manipulation;" the same can be said for "picture," "video," and "generated," suggesting the theme of "AI generated content." At an individual level, the results of KeyBERT showed that many experts think about "defenses" (or "detection" or "countermeasures" or "protection"); intriguingly, the most relevant keyword for one expert, "privacy," had the second-most highest weight among the most relevant keywords for all other experts.

2) ***Open Problems and Concerns of OAI*** *($\mathbb{C}3$):* We use our preliminary analysis as a scaffold and further inspect all expert statements to derive open problems and concerns on OAI. Such a summarization was carried out by four authors who independently formulated their conclusions after reading the statements and then discussed the resulting viewpoints to reach a consensus. While the experts were asked to write statements "describing three open problems of OAI," our analysis showed that some of the statements revealed *specific research problems*, whereas others pertained to more *generic concerns* connected with the OAI threat. Below, we present these two categories of findings separately.

We have identified the following **seven open problems** ($\mathcal{P}$):

$\mathcal{P}1$: *Differentiating AI from Reality.* Content generated by AI has reached quality comparable with reality. This development poses substantial problems for society. Falsification of content by means of AI may have grave consequences for the democratic order, the rule of law, education and numerous other faces of our society. It is of vital importance for the research community to understand the potential effects of such fake audio-visual content and develop corresponding countermeasures.

$\mathcal{P}2$: *Privacy threats of AI.* AI facilitates the extraction of private information about humans, e.g., via attribute inference attacks, linking of separate data items, cross-device tracking, fingerprinting of encrypted traffic. Substantial advances in privacy-related research are needed to counter novel privacy threats enabled by offensive AI.

$\mathcal{P}3$: *Management of offensive AI risks.* The operational implications of OAI in the context of systems cannot be resolved by technical means alone. This puts the problem into a management perspective. To enable decision making, quantification of various risks is required. Attention should be given to the skill level needed for (ab)using certain AI tools, and to the benefits of deployment of AI techniques w.r.t. conventional attacks.

$\mathcal{P}4$: *Implications of offensive AI for social engineering.* Besides a general impact of offensive AI on humans in the societal context, such impact has specific implications for security systems. The risk of humans being the weakest link in a security chain is widely recognized. Offensive AI brings scalability of social engineering attacks to a new level. To counter this threat, both human-centric research (e.g., new methods for awareness training), and system-centric research (e.g., minimizing the likelihood and impact of human error), needs to be pursued.

$\mathcal{P}5$: *Expansion of AI governance.* Offensive AI may damage humans also by impacting specific institutions they interact with. Alongside state regulation (potentially complemented at the international level), new governance mechanisms should be explored to induce/enable institutions to prevent malicious abuse of their AI infrastructures.

$\mathcal{P}6$: *Understanding the pros and cons of offensive AI.* There are a lot of "success stories" about AI-powered attacks. However, not much is known about cases in which such AI-powered attacks resulted in failure—which could be used to shed light on the limitations of AI as an offensive tool. Future work should discuss such negative results as well, which could also entail attacks that are successful, but which are unreasonably expensive to stage in reality.

$\mathcal{P}7$: *Cognitive bias and its implications.* AI can cause cognitive bias in human decisions. An exogenous positive bias induced by offensive AI can strengthen people's existing beliefs and deter critical thinking. AI can also affect behavioral economics, thus eroding economic theories based on the assumption of rational decision making. Understanding of the cognitive impact of offensive AI, as well as of potential "collaboration" between humans and machines to counter such threats, is strongly desirable.

We stress that: *(i)* the above mentioned problems are listed in no specific order of importance; and *(ii)* we do not use "majority" to identify any given problem—even if multiple experts share similar views, every opinion has the same value.

Further, we have identified **three fundamental concerns on OAI** expressed by our experts. Such concerns do not necessarily constitute specific research areas but rather affect the entire realm of research in AI and information security. **(I)** First, *AI is a double-edged sword.* AI was conceived to make human life easier. Unfortunately, it also makes attackers' life easier. Hence the potential dual use of AI must be addressed at various levels, from methodical research to legal regulation, management, and compliance. **(II)** Second, the implications of OAI being able to target *humans who detain different roles and responsibilities*. For instance, tricking an employee has different consequences from tricking a decision maker. Hence, a proper evaluation of OAI threats requires to model all such scenarios and foresee the corresponding effects—which requires assessments of the risks from various viewpoints. **(III)** Third, *countermeasures are needed but challenging to deliver*. On the one hand, there is still little that is known about AI, making it hard to find effective solutions to OAI (and AI being a "double edged sword" further aggravates this challenge); on the other hand, from a research viewpoint, there is a higher incentive in showcasing "novel attacks" rather than on studying, evaluating, and implementing appropriate defenses. Hence, future work should put countermeasures in higher regard—potentially by focusing on techniques that, despite not addressing the OAI threat universally, may just increase the cost to sustain an OAI-based attack.

Remark: We acknowledge that the problems/concerns mentioned above may appear "well-known" in the security community. However, to the best of our knowledge, our SoK is the first scientific work wherein the opinions of 12 experts on OAI have been systematically coalesced into a set of avenues for future research.

C. Comparison of Opinions: Pre and Post Reading this SoK

While summarizing our experts' opinions we have observed some changes in their views which can potentially be attributed to the findings of our paper. Specifically, two topics were stressed more often in the experts' statements (provided after reading our paper) than in the initial responses to the survey.

The first topic is *privacy.* In the initial responses, two experts essentially described OAI use cases that could be related to privacy, but never explicitly associated them with the term "privacy". In written statements, the term "privacy" is explicitly mentioned eight times (by four experts). This may be the result of becoming aware of privacy as an important use-case of OAI, which is revealed in our Fig. 4.

The second topic is *cost.* During our survey, only one expert mentioned "cost" (twice), and the term "econom-" was never mentioned. In contrast, in the statements, three experts mentioned "cost" (the word "cost" occurs 6 times in total), whereas "econom-" was mentioned by four experts (and it occurs 8 times). This fact clearly bears some correlation with the cost/benefit analysis being an essential systematization criterion in our checklist, suggesting that our findings led to an increased awareness on the economical factor of OAI.

VII. Discussion

We now reflect on the findings (§VII-A) and limitations (§VII-B) of our paper and compare it with related meta-research (§VII-C). Our intention is to demonstrate the importance of analyzing various sources of knowledge.

A. Findings

We summarize our findings by making explicit reference to the three-fold contributions of our paper (ℂ1, ℂ2, ℂ3).

Our SoK provides a snapshot of the OAI landscape (ℂ1). Such a snapshot, however, has been made possible only thanks to the collective "contributions" of four sources of knowledge—each of which covers the potential *blind spots* of the others. For instance, our literature review (§III) showed that previous taxonomization of offensive AI use-cases (e.g., reliant on MITRE) are insufficient to cover the landscape of the OAI threat. At the same time, the "limited practicality" exhibited even by technical papers (which mostly attack "toy" systems) may suggest that OAI does not represent a tangible threat—but, perhaps unfortunately, the analysis of the InfoSec briefings (§IV) revealed that OAI can be practically exploited in the real world. In contrast, no InfoSec briefing considered OAI in warfare, but we found many papers covering such use cases—which also seem to worry non-experts (§V), despite not having been mentioned by any of our experts (§VI). Finally, despite our extensive analyses, we acknowledge that our review of prior work may have missed some OAI use cases: one such example are website fingerprinting [210] attacks, which we overlooked in Table I, but which were mentioned by one expert in their statements. Altogether, these findings show that there is a need of a perpetual and collective effort to monitor the threat of OAI, since we expect more OAI use cases (existing or new) to be identified in the future (see, e.g., [211]).

Our SoK provides a foundation for a long-term classification of OAI works (ℂ2). It is obvious that the field of potential offensive use-cases for AI is vast. Our simple checklist (§II-E) encapsulates clear criteria that can be used to systematically analyze works on OAI, thereby aligning the corresponding findings to those provided in this paper. Such a checklist represents a methodological stepping stone for keeping track of future discoveries (technical or theoretical) in the OAI context—which serves to identify potential treatments to the threat of OAI. To facilitate the usage of our checklist by downstream research, we have integrated it in an online website [21] in which we *(i)* **maintain a curated and vetted** archive of OAI-related works, and *(ii)* allow interested individuals to add more works to the archive by **submitting "new entries" after applying our checklist**. We have recorded a 60s video (in our repository [54]) showcasing how to use our tool to add new OAI-related works (e.g., the previously mentioned [210]).

Our SoK provides intriguing avenues for future work (ℂ3). The simple characterization of essential features of 133 OAI works and our survey with 549 laypeople enabled to distill many "lessons learned" for future work (scattered through §III–§V), which have been complemented by the shortlist of problems and concerns derived by analyzing the statements of 12 experts (§VI). The latter include: the need for research on countermeasures, the lack of ethical statements and of general understanding of issues related to the dual use of AI, limited focus on societal impact of OAI, the necessity of cost/benefit analysis as part of risk management, and the importance of the "human dimension" in OAI. In some cases, the issues envisioned by the experts (in their statements and survey) align with those that emerged in our user study with non-experts (e.g., deepfakes, cyberattacks, manipulation); however, it is interesting to see that some "educated" laypeople (most of our respondents have degrees, see Appendix B) may have different thoughts (e.g., the above mentioned "warfare"); moreover, some non-experts appear to be more worried of how AI may negatively impact our lives in the general sense, rather than due to an explicit abuse by attackers—which is a valid concern, despite falling outside our scope. Hence, we argue that future work should tackle OAI-related themes by accounting for different perspectives—all being equally important.

B. Limitations

We identify three main limitations that may affect the validity of our findings and discuss them below.

The first limitation pertains to our *search for the literature review and Infosec venues* (§III and §IV). For the latter, we only included BlackHat and DefCon, but there are more InfoSec venues which do accept briefings on AI (e.g., the RSA conference [212]). For the former, our search queries mostly revolved around "offensive AI" and "offensive security" (see §II-A), but there may have been other terms that could have been used to identify works that fall into our definition of OAI; moreover, querying repositories also has limitations on its own [213]. Therefore, the works considered in our SoK (95 research papers and 38 InfoSec briefings) may under-represent the overall number of works on OAI. Our goal, however, was not to attain complete coverage of existing works (which is clearly unfeasible—as we showed in this SoK). Instead, we provide our "snapshot" by analyzing a subset of works drawn from a systematic search of prior work (academic and industrial) and complemented it with the systematic analysis of knowledge distilled from a broad set of sources.

The second limitation entails the *bias in the population of our user studies* (§V and §VI). For the non-expert survey, our sample clearly cannot cover all laypeople in the world. For the expert opinions, we reached out to 12 individuals with different expertise in fields related to OAI. Again, the lack of complete coverage of such opinions should not pose a major threat to the validity of our claims, since we clearly stated that every opinion is equally valid and reported even the concerns shared by few individuals. Nonetheless, carrying out user studies is notoriously difficult (e.g., some top-tier security papers collect the opinion of 10–20 individuals [214–216]).

The third limitation is the potential *subjectivity of qualitative analyses* [217]. Indeed, reviewing each work (paper or briefing), and reviewing the responses of our surveys as well as the expert statements—all these methodical instruments employed in our SoK rely on analyses carried out by its authors. To mitigate this potential shortcoming, we discussed our findings to clarify doubts and to reach a consensus (§II-A and §II-C); and we also relied on well-known practices and technical algorithms (§VI-B1). Moreover, our extensive appendix provides additional information for reproducing most of our results.

C. Comparison with Related Work

We found no prior work that systematically analyzed the theme of offensive AI to the same extent as done in our SoK.

1) ***Prior Work (Surveys/Summarizations) on OAI****:* First, we found no SoK paper that specifically addressed offensive AI. By turning the attention at "literature reviews" (or similar papers), we found that most such papers considered a *single target* (e.g., only "organizations" [20]; or only "humans" [218], or only "systems" [19]). A notable exception is the recent paper by Malatji and Tolah [219], which accounts for offensive AI from a socio-technical perspective. Yet, the analysis in [219] (as well as those in [19, 218]) is *rooted only on the findings of academic literature*. Remarkably, the work by Mirsky et al. [20] also accounts for the perspective of practitioners, but it does not account for industrial venues, nor investigate the opinion of laypeople—which are among the primary targets of OAI and represent a valuable source of knowledge to pinpoint some of the most perceived concerns. Such a narrower scope may lead to some oversights.

2) ***Novel Findings****:* Our SoK underscored OAI use-cases that have not been emphasized in prior summaries. For instance, Mirsky et al. focus on MITRE ATT&CK, meaning that anything outside such matrix was outside their scope, e.g., there is no mention of "cyber warfare" or "privacy" in the main body of their paper [20]. In contrast, these terms have been mentioned in, e.g., [19, 219]; however, these works overlooked the potential of "attribute inference attacks" or "fingerprinting," both of which can be perpetrated via AI. However, we reiterate that the reason why we captured these additional use cases is due to our SoK having a broader scope. Hence, our SoK extends all such prior work by providing a systematization of *all potential targets* of OAI by accounting for *diverse knowledge sources* (see Fig. 1). Moreover, we also claim that our approach is unprecedented in extant SoK papers.

3) ***Advancing the state of the art in SoK****:* We have studied the 270 SoK papers listed in Shujun Li's online bibliography of SoK papers [220], from 2010 to Jan. 2024. We could not find any paper that considers expert opinion as one of the knowledge sources and presents verbatim such expert opinions. The majority of prior SoK papers draw their conclusions from the scientific literature. Some also carry out user studies (either with experts [32], or among the general public [52]). Yet, we found no SoK that considered both of these dimensions—and, specifically, no SoK paper that reached out to experts in an attempt to *draw avenues for future research*. Hence, our contribution can be inspiring also for future SoK papers. Importantly, some SoKs carry out technical experiments (e.g., by reproducing prior work, such as [221]): these SoKs are orthogonal to ours. However, future "technical SoK" can also benefit from our intuitions, e.g., by systematizing the techniques proposed in InfoSec venues.

VIII. Conclusions

We consider this paper as a first step in laying down a scientific groundwork for investigating various facets of offensive AI.

In short, we found that the offensive capabilities of AI are very heterogeneous and can adversely affect systems, humans and the society as a whole. Due to this heterogeneity, offensive AI use cases cannot be classified into a single framework, such as MITRE ATT&CK, but require a broader systematization which we provide with the help of our OAI assessment checklist (§II-E) and in our online tool [21].

We hope that the insights obtained in this SoK paper enable security and privacy researchers to better appreciate the societal impact of problems related to offensive AI. Our findings also underscore the necessity of interdisciplinary collaboration with the areas of cognitive science, psychology, economics, political science, law, ethics, and perhaps many other, to fully comprehend and mitigate the offensive potential of AI.

Acknowledgement. The authors would like to thank the anonymous reviewers for their feedback. We also extend our gratitude to all participants of our non-expert user survey. Part of this research has been funded by the Hilti Foundation. Moreover, we acknowledge the financial support from the projects SERICS (PE00000014) and FAIR (PE00000013) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. This research has also been partially funded by the Research Fund KU Leuven, and by the Cybersecurity Research Program Flanders.

References

[1] R. Girasa, *Artificial Intelligence as a Disruptive Technology: Economic Transformation and Government Regulation*. Springer, 2020.

[2] V.-D. Păvăloaia and S.-C. Necula, "Artificial intelligence as a disruptive technology—a systematic literature review," *Electronics*, 2023.

[3] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cognitive Robotics*, 2023.

[4] J. P. Bharadiya, "Machine learning and ai in business intelligence: Trends and opportunities," *International Journal of Computer*, 2023.

[5] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios, "What factors contribute to the acceptance of artificial intelligence? a systematic review," *Telematics and Informatics*, 2023.

[6] G. Apruzzese, P. Laskov, E. Montes de Oca, W. Mallouli, L. Brdalo Rapa, A. V. Grammatopoulos, and F. Di Franco, "The role of machine learning in cybersecurity," *ACM Digital Threats: Research and Practice*, 2023.

[7] K. M. Sayler, "Artificial intelligence and national security," *Congressional Research Service*, 2020.

[8] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nature Machine Intelligence*, 2019.

[9] T. Van Ede, H. Aghakhani, N. Spahn, R. Bortolameotti, M. Cova, A. Continella, M. van Steen, A. Peter, C. Kruegel, and G. Vigna, "Deepcase: Semi-supervised contextual analysis of security events," in *IEEE Symposium on Security and Privacy*, 2022.

[10] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, 2018.

[11] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "Sok: Security and privacy in machine learning," in *IEEE European Symposium on Security and Privacy*, 2018.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017.

[13] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[14] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *USENIX Security Symposium*, 2016.

[15] T. Langford and B. Payne, "Phishing faster: Implementing chatgpt into phishing campaigns," in *Future Technologies Conference*, 2023.

[16] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," *IEEE Access*, 2024.

[17] V. Bob, "Ai tools such as chatgpt are generating a mammoth increase in malicious phishing emails," *CNBC*, 2023. [Online]. Available: https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html

[18] Slashnext, "The state of phishing," https://slashnext.com/wp-content/uploads/2023/10/SlashNext-The-State-of-Phishing-Report-2023.pdf, Slashnext, Tech. Rep., 2023.

[19] I. D. Aiyanyo, H. Samuel, and H. Lim, "A systematic review of defensive and offensive cybersecurity with machine learning," *Applied Sciences*, 2020.

[20] Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, M. Pintor, W. Lee, Y. Elovici *et al.*, "The Threat of Offensive AI to Organizations," *Computers & Security*, 2023.

[21] "Long-term Collection (and classification) of works on Offensive AI (website of this paper)," https://sok-offensive-ai.github.io/, 2024.

[22] N. Z. Gong and B. Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," in *USENIX Security Symposium*, 2016.

[23] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *IEEE International Conference on Privacy, Security, Risk and Trust & IEEE International Conference on Social Computing*, 2011.

[24] A. Lohn and M. Musser, "Disinformation at scale: Using gpt-3 maliciously for information operations," in *BlackHat*, 2021.

[25] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on GAN," in *International Conference on Data Mining and Big Data*, 2022.

[26] Z. Lin, Y. Shi, and Z. Xue, "Idsgan: Generative adversarial networks for attack generation against intrusion detection," in *Pacific-asia Conference on Knowledge Discovery and Data Mining*, 2022.

[27] N. Šrndić and P. Laskov, "Practical evasion of a learning-based classifier: A case study," in *IEEE Symposium on Security and Privacy*, 2014.

[28] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, 2015.

[29] N. Kaloudi and J. Li, "The AI-based cyber threat landscape: A survey," *ACM Computing Surveys*, 2020.

[30] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The emerging threat of ai-driven cyber attacks: A review," *Applied Artificial Intelligence*, 2022.

[31] "Guidelines for performing systematic literature reviews in software engineering," School of Computer Science and Mathematics, Keele University & Department of Computer Science, University of Durham, Tech. Rep., 2007.

[32] P. Ladisa, H. Plate, M. Martinez, and O. Barais, "Sok: Taxonomy of attacks on open-source software supply chains," in *IEEE Symposium on Security and Privacy*, 2023.

[33] L. Erdődi, Å. Å. Sommervoll, and F. M. Zennaro, "Simulating SQL injection vulnerability exploitation using Q-learning reinforcement learning agents," *Journal of Information Security and Applications*, 2021.

[34] F. M. Zennaro and L. Erdődi, "Modelling penetration testing with reinforcement learning using capture-the-flag challenges: Trade-offs between model-free learning and a priori knowledge," *IET Information Security*, 2023.

[35] O. Valea and C. Oprişa, "Towards pentesting automation using the metasploit framework," in *IEEE International Conference on Intelligent Computer Communication and Processing*, 2020.

[36] M. Ceccato, C. D. Nguyen, D. Appelt, and L. C. Briand, "SOFIA: An automated security oracle for black-box testing of SQL-injection vulnerabilities," in *IEEE/ACM International Conference on Automated Software Engineering*, 2016.

[37] A. Bhattacharya, T. Ramachandran, S. Banik, C. P. Dowling, and S. D. Bopardikar, "Automated adversary emulation for cyber-physical systems via reinforcement learning," in *IEEE International Conference on Intelligence and Security Informatics*, 2020.

[38] W. R. Hersh, A. M. Totten, K. B. Eden, B. Devine, P. Gorman, S. Z. Kassakian, S. S. Woods, M. Daeges, M. Pappas, and M. S. McDonagh, "Outcomes from health information exchange: systematic review and future research needs," *JMIR Medical Informatics*, 2015.

[39] I. Jana and A. Oprea, "Appmine: Behavioral analytics for web application vulnerability detection," in *ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2019.

[40] G. Apruzzese, M. Conti, and Y. Yuan, "Spacephish: the evasion-space of adversarial attacks against phishing website detectors using machine learning," in *Annual Computer Security Applications Conference*, 2022.

[41] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[42] D. R. McKinnel, T. Dargahi, A. Dehghantanha, and K.-K. R. Choo, "A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment," *Computers & Electrical Engineering*, 2019.

[43] O. Kubovič, P. Košinár, and J. Jánošík, "Can artificial intelligence power future malware," *ESET white paper*, 2018.

[44] Forbes, https://www.forbes.com/sites/forbestechcouncil/2023/07/17/mitigating-ai-based-cyberattacks/, 2023.

[45] Economist, https://www.economist.com/biometrics-pod, 2023.

[46] CNN, https://edition.cnn.com/videos/business/2023/05/19/exp-ai-signifyd-intv-051909aseg1-cnni-business.cnn.

[47] https://www.linkedin.com/posts/ben-nassi-phd-68a743115_i-have-recently-received-a-few-emails-from-activity-7161264634505703424--NYk?utm_source=share&utm_medium=member_deskto, 2024.

[48] J. Wilkin, https://www.blackhat.com/us-18/arsenal.html#jacob-wilkin, 2018.

[49] P. Golle, "Machine learning attacks against the Asirra CAPTCHA," in *ACM SIGSAC Conference on Computer and Communications Security*, 2008.

[50] I. Clarke, "Hacking desire: Reverse-engineering what people wan," in *DefCon*, 2008.

[51] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar *et al.*, "Sok: Hate, harassment, and the changing landscape of online abuse," in *IEEE Symposium on Security and Privacy*, 2021.

[52] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee, "Sok: Authentication in augmented and virtual reality," in *IEEE Symposium on Security and Privacy*, 2022.

[53] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The Menlo report," *IEEE Security & Privacy*, 2012.

[54] "Our repository," https://github.com/hihey54/sok_oai/, 2024.

[55] R. W. Emerson, "Convenience sampling, random sampling, and snow-

ball sampling: How does sampling affect the validity of research?” *Journal of Visual Impairment & Blindness*, 2015.

[56] C. Antoun, C. Zhang, F. G. Conrad, and M. F. Schober, “Comparisons of online recruitment strategies for convenience samples: Craigslist, google adwords, facebook, and amazon mechanical turk,” *Field methods*, 2016.

[57] K. Charmaz, *Constructing grounded theory: A practical guide through qualitative analysis*. SAGE Publications Ltd, 2006.

[58] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, and A. Jeyaraj et al., ““So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *International Journal of Information Management*, 2023.

[59] A. Clark, “Whatever next? predictive brains, situated agents, and the future of cognitive science,” *Behavioral and brain sciences*, 2013.

[60] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, and M. Bennis et al., “Advances and open problems in federated learning,” *Foundations and Trends in Machine Learning*, 2021.

[61] R. J. Chenail, “Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research.” *Qualitative report*, 2011.

[62] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, ““Real Attackers Don’t Compute Gradients”: Bridging the Gap Between Adversarial ML Research and Practice,” in *IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[63] https://attack.mitre.org/.

[64] https://attack.mitre.org/matrices/enterprise/.

[65] https://attack.mitre.org/matrices/mobile/.

[66] https://attack.mitre.org/matrices/ics/.

[67] P. P. Tricomi, L. Facciolo, G. Apruzzese, and M. Conti, “Attribute inference attacks in online multiplayer video games: A case study on Dota2,” in *ACM Conference on Data and Application Security and Privacy*, 2023.

[68] N. Yu and K. Darling, “A low-cost approach to crack python captchas using ai-based chosen-plaintext attack,” *Applied Sciences*, 2019.

[69] A. Draganovic, S. Dambra, J. A. Iuit, K. Roundy, and G. Apruzzese, “”Do users fall for real adversarial phishing?” Investigating the human response to evasive webpages,” in *APWG Symposium on Electronic Crime Research*, 2023.

[70] K. Xu, Y. Guo, L. Guo, Y. Fang, and X. Li, “My privacy my decision: Control of photo sharing on online social networks,” *IEEE Transactions on Dependable and Secure Computing*, 2015.

[71] G. Sampagnaro, “Keyword occurrences and journal specialization,” *Scientometrics*, 2023.

[72] D. Antonelli, R. Cascella, A. Schiano, G. Perrone, and S. P. Romano, ““dirclustering”: a semantic clustering approach to optimize website structure discovery during penetration testing,” *Journal of Computer Virology and Hacking Techniques*, 2024.

[73] A. AlMajali, L. Al-Abed, R. Mutleq, Z. Samamah, A. A. Shhadeh, B. J. Mohd, and K. M. A. Yousef, “Vulnerability exploitation using reinforcement learning,” in *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology*, 2023.

[74] J. Chen, S. Hu, H. Zheng, C. Xing, and G. Zhang, “GAIL-PT: An intelligent penetration testing framework with generative adversarial imitation learning,” *Computers & Security*, 2023.

[75] A. Chowdhary, K. Jha, and M. Zhao, “Generative Adversarial Network (GAN)-Based Autonomous Penetration Testing for Web Applications,” *Sensors*, 2023.

[76] P. Gallus, M. Štěpánek, T. Ráčil, and P. Františ, “Generative neural networks as a tool for web applications penetration testing,” in *IEEE Communication and Information Technologies*, 2023.

[77] M. C. Ghanem, T. M. Chen, and E. G. Nepomuceno, “Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks,” *Journal of Intelligent Information Systems*, 2023.

[78] A. Happe and J. Cito, “Getting pwn’d by ai: Penetration testing with large language models,” in *ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023.

[79] F. Iqbal, F. Samsom, F. Kamoun, and Á. MacDermott, “When ChatGPT goes rogue: exploring the potential cybersecurity threats of AI-powered conversational chatbots,” *Frontiers in Communications and Networks*, 2023.

[80] E. Karinshak, S. X. Liu, J. S. Park, and J. T. Hancock, “Working with ai to persuade: Examining a large language model’s ability to generate pro-vaccination messages,” *Proceedings of the ACM on Human-Computer Interaction*, 2023.

[81] O. S. Ozturk, E. Ekmekcioglu, O. Cetin, B. Arief, and J. Hernandez-Castro, “New tricks to old codes: can ai chatbots replace static code analysis tools?” in *European Interdisciplinary Cybersecurity Conference*, 2023.

[82] Y. M. Pa Pa, S. Tanizaki, T. Kou, M. Van Eeten, K. Yoshioka, and T. Matsumoto, “An attacker’s dream? exploring the capabilities of chatgpt for developing malware,” in *Cyber Security Experimentation and Test Workshop*, 2023.

[83] N. Auricchio, A. Cappuccio, F. Caturano, G. Perrone, and S. P. Romano, “An automated approach to web offensive security,” *Computer Communications*, 2022.

[84] D. Biesner, K. Cvejoski, and R. Sifa, “Combining variational autoencoders and transformer language models for improved password generation,” in *International Conference on Availability, Reliability and Security*, 2022.

[85] T. Cody, A. Rahman, C. Redino, L. Huang, R. Clark, A. Kakkar, D. Kushwaha, P. Park, P. Beling, and E. Bowen, “Discovering exfiltration paths using reinforcement learning with attack graphs,” in *IEEE Conference on Dependable and Secure Computing*, 2022.

[86] A. Confido, E. V. Ntagiou, and M. Wallum, “Reinforcing Penetration Testing Using AI,” in *IEEE Aerospace Conference*, 2022.

[87] R. Gangupantulu, T. Cody, P. Park, A. Rahman, L. Eisenbeiser, D. Radke, R. Clark, and C. Redino, “Using cyber terrain in reinforcement learning for penetration testing,” in *IEEE International Conference on Omni-layer Intelligent Systems*, 2022.

[88] R. S. Jagamogan, S. A. Ismail, N. H. Hassan, and H. Abas, “Penetration testing procedure using machine learning,” in *International Conference on Smart Sensors and Application*, 2022.

[89] D. Karanatsiou, P. Sermpezis, D. Gruda, K. Kafetsios, I. Dimitriadis, and A. Vakali, “My tweets bring all the traits to the yard: Predicting personality and relational traits in online social networks,” *ACM Transactions on the Web*, 2022.

[90] S. Lee, S. Wi, and S. Son, “Link: Black-box detection of cross-site scripting vulnerabilities using reinforcement learning,” in *The ACM Web Conference 2022*, 2022.

[91] Y. Li, J. Yan, and M. Naili, “Deep reinforcement learning for penetration testing of cyber-physical attacks in the smart grid,” in *International Joint Conference on Neural Networks*, 2022.

[92] N. X. Nhu, T. T. Nghia, N. H. Quyen, V.-H. Pham, P. T. Duy *et al.*, “Leveraging deep reinforcement learning for automating penetration testing in reconnaissance and exploitation phase,” in *IEEE International Conference on Computing and Communication Technologies*, 2022.

[93] G. Pagnotta, D. Hitaj, F. De Gaspari, and L. V. Mancini, “Passflow: guessing passwords with generative flows,” in *Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2022.

[94] K. Tran, M. Standen, J. Kim, D. Bowman, T. Richer, A. Akella, and C.-T. Lin, “Cascaded reinforcement learning agents for large action spaces in autonomous penetration testing,” *Applied Sciences*, 2022.

[95] Q. Yao, Y. Wang, X. Xiong, and Y. Li, “Intelligent penetration testing in dynamic defense environment,” in *International Conference on Cyber Security*, 2022.

[96] F. Caturano, G. Perrone, and S. P. Romano, “Discovering reflected cross-site scripting vulnerabilities using a multiobjective reinforcement learning environment,” *Computers & Security*, 2021.

[97] R. Gangupantulu, T. Cody, A. Rahma, C. Redino, R. Clark, and P. Park, “Crown jewels analysis using reinforcement learning with attack graphs,” in *IEEE Symposium Series on Computational Intelligence*, 2021.

[98] H. Khan, M. Alam, S. Al-Kuwari, and Y. Faheem, “Offensive AI: Unification of email generation through GPT-2 with a game-theoretic approach for spear-phishing attacks,” *Competitive Advantage in the Digital Economy*, 2021.

[99] K. Kujanpää, W. Victor, and A. Ilin, “Automating privilege escalation with deep reinforcement learning,” in *ACM Workshop on Artificial Intelligence and Security*, 2021.

[100] K. Lee, J. Lee, C. Choi, and K. Yim, “Offensive security of keyboard data using machine learning for password authentication in iot,” *IEEE Access*, 2021.

[101] R. Maeda and M. Mimura, “Automating post-exploitation with deep reinforcement learning,” *Computers & Security*, 2021.

[102] C. Neal, H. Dagdougui, A. Lodi, and J. M. Fernandez, “Reinforcement

learning based penetration testing of a microgrid control algorithm," in *IEEE Annual Computing and Communication Workshop and Conference*, 2021.

[103] F. Sharevski, P. Jachim, and E. Pieroni, "Regulation tl; dr: Adversarial text summarization of federal register articles," in *Workshop on Cyber-Security Arms Race*, 2021.

[104] M. Standen, M. Lucas, D. Bowman, T. Richer, J. Kim, and D. Marriott, "CybORG: A Gym for the Development of Autonomous Cyber Agents," in *International Workshop on Adaptive Cyber Defense (co-located with IJCAI)*, 2021.

[105] C. Toemmel, "Catch Me If You GAN: Using Artificial Intelligence for Fake Log Generation," *arXiv:2112.12006*, 2021.

[106] K. Tran, A. Akella, M. Standen, J. Kim, D. Bowman, T. Richer, and C.-T. Lin, "Deep hierarchical reinforcement agents for automated penetration testing," *arXiv:2109.06449*, 2021.

[107] A. Al-Hababi and S. C. Tokgoz, "Man-in-the-middle attacks to detect and identify services in encrypted network flows using machine learning," in *IEEE International Conference on Advanced Communication Technologies and Networking*, 2020.

[108] A. Chowdhary, D. Huang, J. S. Mahendran, D. Romo, Y. Deng, and A. Sabur, "Autonomous security analysis and penetration testing," in *International Conference on Mobility, Sensing and Networking*, 2020.

[109] A. Halimi and E. Ayday, "Efficient quantification of profile matching risk in social networks using belief propagation," in *European Symposium on Research in Computer Security*, 2020.

[110] Z. Hu, R. Beuran, and Y. Tan, "Automated penetration testing using deep reinforcement learning," in *IEEE European Symposium on Security and Privacy Workshops*, 2020.

[111] K. Lee and K. Yim, "Cybersecurity threats based on machine learning-based offensive technique for password authentication," *Applied Sciences*, 2020.

[112] K. Lee and S.-Y. Lee, "Improved practical vulnerability analysis of mouse data according to offensive security based on machine learning in image-based user authentication," *Entropy*, 2020.

[113] M. Liu, K. Li, and T. Chen, "Deepsqli: Deep semantic learning for testing sql injection," in *ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2020.

[114] W. Pearce, N. Landers, and N. Fulda, "Machine learning for offensive security: sandbox classification using decision trees and artificial neural networks," in *Computing Conference*, 2020.

[115] F. Sharevski, P. Jachim, and E. Pieroni, "Wikipediabot: Machine learning assisted adversarial manipulation of wikipedia articles," in *Workshop on DYnamic and Novel Advances in Machine Learning and Intelligent Cyber Security*, 2020.

[116] D. Shu, N. O. Leslie, C. A. Kamhoua, and C. S. Tucker, "Generative adversarial attacks against intrusion detection systems using active learning," in *ACM Workshop on Wireless Security and Machine Learning*, 2020.

[117] W. Song, X. Li, S. Afroz, D. Garg, D. Kuznetsov, and H. Yin, "Mab-malware: A reinforcement learning framework for blackbox generation of adversarial malware," in *ACM Asia Conference on Computer and Communications Security*, 2022.

[118] N. Yu, Z. Tuttle, C. J. Thurnau, and E. Mireku, "Ai-powered gui attack and its defensive methods," in *ACM Southeast Conference*, 2020.

[119] C. Basu, S. Venkatesan, C.-Y. J. Chiang, N. Leslie, and C. Kamhoua, "Generating targeted e-mail at scale using neural machine translation," in *Workshop on DYnamic and Novel Advances in Machine Learning and Intelligent Cyber Security*, 2019.

[120] S. Cecconello, A. Compagno, M. Conti, D. Lain, and G. Tsudik, "Skype & type: Keyboard eavesdropping in voice-over-ip," *ACM Transactions on Privacy and Security*, 2019.

[121] K. Chung, Z. T. Kalbarczyk, and R. K. Iyer, "Availability attacks on computing systems through alteration of environmental control: smart malware approach," in *ACM/IEEE International Conference on Cyber-Physical Systems*, 2019.

[122] D. Das, A. Golder, J. Danial, S. Ghosh, A. Raychowdhury, and S. Sen, "X-DeepSCA: Cross-device deep learning side channel attack," in *Annual Design Automation Conference*, 2019.

[123] M. C. Ghanem and T. M. Chen, "Reinforcement learning for efficient network penetration testing," *Information*, 2019.

[124] J. M. Tshimula, B. Chikhaoui, and S. Wang, "Har-search: A method to discover hidden affinity relationships in online communities," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019.

[125] R. Zhang, X. Chen, S. Wen, X. Zheng, and Y. Ding, "Using ai to attack va: a stealthy spyware against voice assistances in smart phones," *IEEE Access*, 2019.

[126] S. A. Anand and N. Saxena, "Keyboard emanations in remote voice calls: Password leakage and noise (less) masking defenses," in *ACM Conference on Data and Application Security and Privacy*, 2018.

[127] A. C. Bahnsen, I. Torroledo, L. D. Camacho, and S. Villegas, "Deep-phish: simulating malicious ai," in *APWG Symposium on Electronic Crime Research*, 2018.

[128] J. Kronjee, A. Hommersom, and H. Vranken, "Discovering software vulnerabilities using data-flow analysis and machine learning," in *International Conference on Availability, Reliability and Security*, 2018.

[129] M. Rigaki and S. Garcia, "Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection," in *IEEE Security and Privacy Workshops*, 2018.

[130] F. Zhou, L. Liu, K. Zhang, G. Trajcevski, J. Wu, and T. Zhong, "Deeplink: A deep learning approach for user identity linkage," in *IEEE Conference on Computer Communications*, 2018.

[131] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *ACM SIGSAC Conference on Computer and Communications Security*, 2017.

[132] H. S. Anderson, J. Woodbridge, and B. Filar, "DeepDGA: Adversarially-tuned domain generation and detection," in *ACM Workshop on Artificial Intelligence and Decurity*, 2016.

[133] G. Grieco, G. L. Grinblat, L. Uzal, S. Rawat, J. Feist, and L. Mounier, "Toward large-scale vulnerability discovery using machine learning," in *ACM Conference on Data and Application Security and Privacy*, 2016.

[134] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, "Reverse engineering socialbot infiltration strategies in twitter," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2015.

[135] E. Bursztein, J. Aigrain, A. Moscicki, and J. C. Mitchell, "The end is nigh: Generic solving of text-based CAPTCHAs," in *USENIX Workshop on Offensive Technologies*, 2014.

[136] S. Adali and J. Golbeck, "Predicting personality with social behavior," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[137] A. Malhotra, L. Totti, W. Meira Jr, P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[138] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets," in *International Conference on Machine Learning and Applications*, 2012.

[139] F. Yamaguchi, K. Rieck *et al.*, "Vulnerability extrapolation: Assisted discovery of vulnerabilities using machine learning," in *USENIX Workshop on Offensive Technologies*, 2011.

[140] E. Bursztein and S. Bethard, "Decaptcha: breaking 75% of ebay audio captchas," in *USENIX Conference on Offensive technologies*, 2009.

[141] J. Kelly, M. DeLaus, E. Hemberg, and U.-M. O'Reilly, "Adversarially adapting deceptive views and reconnaissance scans on a software defined network," in *IFIP/IEEE Symposium on Integrated Network and Service Management*, 2019.

[142] Future of Life, "Autonomous Weapons Open Letter: AI & Robotics Researchers," https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/, 2016.

[143] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan, "Detecting algorithmically generated domain-flux attacks with DNS traffic analysis," *IEEE/ACM Transactions on Networking*, 2012.

[144] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon, "From Throw-Away traffic to bots: Detecting the rise of DGA-Based malware," in *USENIX Security Symposium*, 2012.

[145] A. C. Dall'Agnol, "Artificial intelligence and the future of warfare: The usa, china, and strategic stability," *Journal of Strategic Studies*, 2023.

[146] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, and C. Rizzo, "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health," *Frontiers in Public Health*, 2023.

[147] O. Illiashenko, V. Kharchenko, I. Babeshko, H. Fesenko, and F. Di Giandomenico, "Security-informed safety analysis of autonomous transport systems considering ai-powered cyberattacks and protection,"

Entropy, 2023.

[148] E. Pashentsev, “Destabilization of unstable dynamic social equilibriums and the malicious use of artificial intelligence in high-tech strategic psychological warfare,” in *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Springer, 2023.

[149] J.-M. Rickli and F. Mantellassi, “Artificial intelligence in warfare: Military uses of ai and their international security implications,” in *The AI Wave in Defence Innovation*. Routledge, 2023.

[150] W. Hao, C. Shen, X. Yang, and C. Wang, “Intelligent penetration and attack simulation system based on attack chain,” in *International Symposium on Computational Intelligence and Design*, 2022.

[151] S. Kasim, N. Valliani, N. K. K. Wong, S. Samadi, L. Watkins, and A. Rubin, “Cybersecurity as a Tic-Tac-Toe Game Using Autonomous Forwards (Attacking) And Backwards (Defending) Penetration Testing in a Cyber Adversarial Artificial Intelligence System,” in *IEEE International Conference of Computer Science and Information Technology*, 2022.

[152] R. McIlroy-Young, J. Kleinberg, S. Sen, S. Barocas, and A. Anderson, “Mimetic models: Ethical implications of ai that acts like you,” in *AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

[153] C. Nica and T. Tănase, “Using weaponized machine learning in cyber offensive operations,” in *International Conference: The Knowledge-based Organization*, 2020.

[154] P. Skeba and E. P. Baumer, “Informational friction as a lens for studying algorithmic aspects of privacy,” *Proceedings of the ACM on Human-Computer Interaction*, 2020.

[155] C. Easttom, “A methodological approach to weaponizing machine learning,” in *International Conference on Artificial Intelligence and Advanced Manufacturing*, 2019.

[156] I. Burton and J. Straub, “Autonomous distributed electronic warfare system of systems,” in *Annual Conference System of Systems Engineering*, 2019.

[157] J. Burton and S. R. Soare, “Understanding the strategic implications of the weaponization of artificial intelligence,” in *International Conference on Cyber Conflict*, 2019.

[158] A. Giaretta and N. Dragoni, “Community targeted phishing: A middle ground between massive and spear phishing through natural language generation,” in *International Conference in Software Engineering for Defence Applications*, 2019.

[159] G. Maus, “Decoding, hacking, and optimizing societies: Exploring potential applications of human data analytics in sociological engineering, both internally and as offensive weapons,” in *IEEE Science and Information Conference*, 2015.

[160] A. Guarino, “Autonomous intelligent agents in cyber offence,” in *International Conference on Cyber Conflict*, 2013.

[161] D. Olszewski, A. Lu, C. Stillman, K. Warren, C. Kitroser, A. Pascual, D. Ukirde, K. Butler, and P. Traynor, “”Get in Researchers; We’re Measuring Reproducibility”: A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences,” in *ACM SIGSAC Conference on Computer and Communications Security*, 2023.

[162] E. Scheiner and B. Bivu, “Synthetic trust: Exploiting biases at scale,” in *BlackHat*, 2023.

[163] M. Canham and B. Sawyer, “Me and my evil digital twin: The psychology of human exploitation by ai assistants,” in *BlackHat*, 2023.

[164] F. Heiding, B. Schneier, A. Vishwanath, and J. Bernstein, “Devising and Detecting Phishing: Large Language Models (GPT3, GPT4) vs. Smaller Human Models (V-Triad, Generic Emails),” in *BlackHat*, 2023.

[165] ——, “Devising and detecting phishing: Large language models vs. smaller human models,” in *BlackHat*, 2023.

[166] A. Herbert-Voss and S. Caldwell, “How NOT to Train Your Hack Bot: Dos and Dont’s of Building Offensive GPTs,” in *BlackHat*, 2023.

[167] W. Waligóra, “How we taught ChatGPT-4 to break Mbed TLS AES with side-channel attacks,” in *BlackHat*, 2023.

[168] C. Gibson, V. Kropotov, and F. Yarochkin, “Kidnapping without hostages: Virtual kidnapping and the dark road ahead,” in *BlackHat*, 2023.

[169] G. Zror, “Look ma i’m the ceo! real-time video and audio deep-fake!” in *DefCon*, 2023.

[170] L. Xin and T. Yuan, “Human or not: Can you really detect the fake voices?” in *BlackHat*, 2022.

[171] C. Chi, “Bridging the gap between research and practice in intelligently bypassing waf,” in *BlackHat*, 2022.

[172] E. Lim, G. Tan, T. K. Hock, and T. Lee, “Hacking humans with ai as a service,” in *DefCon*, 2021.

[173] A. Lohn and M. Musser, “Disinformation at scale: Using gpt-3 maliciously for information operations,” in *BlackHat*, 2021.

[174] P. Tully and L. Foster, “Repurposing neural networks,” in *BlackHat*, 2020.

[175] T. Basu, “How i clone myself using ai - next gen social engineering,” in *BlackHat*, 2020.

[176] A. Sharma and A. M. Yi, “Effective vulnerability discovery with machine learning,” in *BlackHat*, 2020.

[177] Y. Chen, A. E. Santosa, A. Sharma, and D. Lo, “Automated identification of libraries from vulnerability data,” in *ACM/IEEE International Conference on Software Engineering: Software Engineering in Practice*, 2020.

[178] I. Takaesu, M. Masuya, and T. Yoneyama, “Gyoithon,” in *BlackHat*, 2019.

[179] J. Botwicz, “Cotopaxi iot protocols testing toolkit,” in *DefCon*, 2019.

[180] E. Bursztein, “Deep learning revolutionizing side channel cryptanalysis,” in *DefCon*, 2019.

[181] L. Ding, A. Benameur, J. Jacob, J. Chen, and S. Pham, “Automated rest api endpoint identification for security testing at scale,” in *BlackHat*, 2019.

[182] M. Price and M. Price, “Playing offense and defense with deepfakes,” in *BlackHat*, 2019.

[183] A. C. Bahnsen, “Deepphish simulating malicious ai,” in *BlackHat*, 2018.

[184] R. Greenstadt and A. Caliskan, “De-anonymizing programmers from source code and binaries,” in *DefCon*, 2018.

[185] A. Caliskan, F. Yamaguchi, E. Dauber, R. Harang, K. Rieck, R. Greenstadt, and A. Narayanan, “When coding style survives compilation: De-anonymizing programmers from executable binaries,” *arXiv:1512.08546*, 2015.

[186] D. Kirat, J. Jang, and M. P. Stoecklin, “Deeplocker: Concealing targeted attacks with ai locksmithing,” in *BlackHat*, 2018.

[187] G. Perin, B. Ege, and J. V. Woudenberg, “Lowering the bar: Deep learning for side channel analysis,” in *BlackHat*, 2018.

[188] F. Gomez and C. Jimenez, “Video killed the text star: Osint approach,” in *BlackHat*, 2018.

[189] H. Anderson, “Evading next-gen av using a.i.” in *DefCon*, 2017.

[190] H. S. Anderson, A. Kharkar, B. Filar, D. Evans, and P. Roth, “Learning to evade static PE machine learning malware models via reinforcement learning,” *arXiv:1801.08917*, 2018.

[191] D. Lain, M. Conti, G. Tsudik, and A. Compagno, “Skype & type: Keystroke leakage over voip,” in *BlackHat*, 2017.

[192] B. Morris and D. Petro, “Weaponizing machine learning: Humanity was overrated anyway,” in *DefCon*, 2017.

[193] P. Tully and M. Raggo, “A picture is worth a thousand words, literally: Deep neural networks for social stego,” in *DefCon*, 2017.

[194] A. Singh and V. Thaware, “Wire me through machine learning,” in *BlackHat*, 2017.

[195] I. Polakis and S. Sivakorn, “I’m not a human: Breaking the google recaptcha,” in *BlackHat*, 2016.

[196] S. Sivakorn, I. Polakis, and A. D. Keromytis, “I am robot:(deep) learning to break semantic image captchas,” in *IEEE European Symposium on Security and Privacy*, 2016.

[197] G. Argyros and I. Stais, “Another brick off the wall: Deconstructing web application firewalls using automata learning,” in *BlackHat*, 2016.

[198] G. Argyros, I. Stais, S. Jana, A. D. Keromytis, and A. Kiayias, “Sfadiff: Automated evasion attacks and fingerprinting using black-box differential automata learning,” in *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[199] J. Seymour and P. Tully, “Weaponizing data science for social engineering: Automated e2e spear phishing on twitter,” in *DefCon*, 2016.

[200] M. Wolff, B. Wallace, S. Researcher, and X. Zhao, “Applied machine learning for data exfiltration and other fun topics,” in *BlackHat*, 2016.

[201] C. Bursztein and E. Bursztein, “I am a legend: Hacking hearthstone with machine learning,” in *DefCon*, 2014.

[202] X. Fu, Q. Yue, and Z. Ling, “My google glass sees your passwords!” in *BlackHat*, 2014.

[203] Q. Yue, Z. Ling, X. Fu, B. Liu, K. Ren, and W. Zhao, “Blind recognition of touched keys on mobile devices,” in *ACM SIGSAC Conference on Computer and Communications Security*, 2014.

[204] S. Vanned, “Evolving exploits through genetic algorithms,” in *DefCon*, 2013.

[205] J. Espinhara and U. Albuquerque, “Using online activity as digital fingerprints to create a better spear phisher,” in *BlackHat*, 2013.

[206] I. Clarke, "Hacking desire: Reverse-engineering what people want," in *DefCon*, 2008.

[207] "Global cybersecurity outlook," https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2024.pdf, World Economic Forum, Tech. Rep., 2024.

[208] M. Grootendorst, "Keybert: Minimal keyword extraction with bert." 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265

[209] ——, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv:2203.05794*, 2022.

[210] V. Rimmer, D. Preuveneers, M. Juarez, T. Van Goethem, and W. Joosen, "Automated website fingerprinting through deep learning," in *Network and Distributed Systems Security Symposium*, 2018.

[211] A. Dragan, H. King, and A. Dafoe, "Frontier safety framework (v1.0)," Google DeepMind, Tech. Rep., 2024. [Online]. Available: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf

[212] SiliconAngle, "RSA Conference 2024 goes beyond AI-powered security to securing AI itself," https://web.archive.org/web/20240601053332/https://siliconangle.com/2024/05/11/rsa-conference-2024-goes-beyond-ai-powered-security-securing-ai/, 2024.

[213] W. M. Bramer, "Variation in number of hits for complex searches in google scholar," *Journal of the Medical Library Association*, 2016.

[214] B. A. Alahmadi, L. Axon, and I. Martinovic, "99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms," in *USENIX Security Symposium*, 2022.

[215] G. Apruzzese, P. Laskov, and J. Schneider, "SoK: Pragmatic Assessment of Machine Learning for Network Intrusion Fetection," in *IEEE European Symposium on Security and Privacy*, 2023.

[216] J. Mink, H. Benkraouda, L. Yang, A. Ciptadi, A. Ahmadzadeh, D. Votipka, and G. Wang, "Everybody's got ML, tell me what else you have: Practitioners' perception of ML-based security tools and explanations," in *IEEE Symposium on Security and Privacy*, 2023.

[217] A. Madill, A. Jordan, and C. Shirley, "Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies," *British journal of psychology*, 2000.

[218] T. Arora and R. Soni, "A review of techniques to detect the GAN-generated fake images," *Generative Adversarial Networks for Image-to-Image Translation*, 2021.

[219] M. Malatji and A. Tolah, "Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI," *AI and Ethics*, 2024.

[220] S. Li, "Shujun Li's Bibliography of SoK Papers," https://www.hooklee.com/Research/SoK/SoK.html, 2023.

[221] G. Apruzzese, P. Laskov, and A. Tastemirova, "Sok: The impact of unlabelled data in cyberthreat detection," in *IEEE European Symposium on Security and Privacy*, 2022.

[222] D. Nettleton, *Commercial data mining: processing, analysis and modeling for predictive analytics projects*. Elsevier, 2014.

[223] A. Gerace, "Producing fake information is getting easier," *Economist*, 2024. [Online]. Available: https://www.economist.com/science-and-technology/2024/05/01/producing-fake-information-is-getting-easier

[224] "The Cyber Kill Chain," https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

[225] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy*, 2017.

[226] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *ACM SIGSAC Conference on Computer and Communications Security*, 2023.

[227] I. Elsharef, Z. Zeng, and Z. Gu, "Facilitating threat modeling by leveraging large language models," in *Workshop on AI Systems with Confidential Computing*, 2024.

[228] U. H. Govindarajan, D. K. Singh, and H. A. Gohel, "Forecasting cyber security threats landscape and associated technical trends in telehealth using bidirectional encoder representations from transformers (bert)," *Computers & Security*, 2023.

[229] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, "C-pack: Packaged resources to advance general chinese embedding," 2023.

[230] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv:2307.09288*, 2023.

[231] IEEE, "Ethical Requirements," https://web.archive.org/web/20241202060241/https://books.ieeeauthorcenter.ieee.org/book-publishing-at-ieee/publishing-ethics/ethical-requirements/, 2024.

[232] Elsevier, "CRediT author statement," https://web.archive.org/web/20241130183511/https://www.elsevier.com/researcher/author/policies-and-guidelines/credit-author-statement, 2024.

[233] "Scimago Journal & Country Rank," https://www.scimagojr.com/.

[234] "Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference," https://dl.acm.org/doi/proceedings/10.1145/3590777.

[235] "Midjourney V3 release," https://en.wikipedia.org/wiki/Midjourney.

[236] "Stable Diffusion release," https://stability.ai/news/stable-diffusion-announcement.

[237] "ChatGPT release," https://openai.com/index/chatgpt/.

[238] "ElevenLabs release," https://elevenlabs.io/blog/elevenlabs-raises-2m-pre-seed-and-announces-ai-speech-platform-promising-to-revolutionize-audio-storytelling.

[239] "Google Bard release," https://blog.google/technology/ai/bard-google-ai-search-updates/.

[240] "Claude launch," https://www.anthropic.com/news/claude-3-family.

[241] "Adobe Firefly launch," https://news.adobe.com/news/2024/10/101424-adobe-launches-firefly-video-model.

[242] "OpenAI DALLE3 launch," https://openai.com/index/new-models-and-developer-products-announced-at-devday/.

[243] "Google Gemini launch," https://blog.google/technology/ai/google-gemini-ai/.

[244] "OpenAI Sora announcement," https://openai.com/index/sora/.

[245] "Mirai launch," https://mistral.ai/news/mixtral-of-experts/.

[246] "ChatGPT-4o launch," https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/.

[247] "GPT-4 launch," https://openai.com/index/gpt-4-research/.

[248] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, "Dos and don'ts of machine learning in computer security," in *USENIX Security Symposium*, 2022.

[249] W. M. Si, M. Backes, J. Blackburn, E. De Cristofaro, G. Stringhini, S. Zannettou, and Y. Zhang, "Why so toxic? measuring and triggering toxic behavior in open-domain chatbots," in *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

Appendix A
Checklist for analyzing OAI-related works

We explain at a low-level the criteria embedded in our proposed offensive AI checklist. We use our checklist to analyze each work considered in this SoK, but our checklist can also be used by future research to carry out analyses that align with ours and expand the findings of our systematization.

A. *What is the OAI use-case?*

First, we map each paper to one of the MITRE ATT&CK Tactics for Enterprise, Mobile or ICS [63].[6] We associate each paper with the primary MITRE Tactic covered: For instance, if the goal of a side-channel attack is password stealing [126], we map it to Credential Access. Some OAI use cases cannot be mapped to the MITRE ATT&CK tactics, as they entail techniques beyond the matrices' scope. These OAI use cases either: encompass multiple tactics—which we classify under the category of autonomous agents; or focus more broadly on attacks against "society" or "privacy," as well as applications in (cyber) "warfare."

Second, we review the original focus of the paper, which we do by asking ourselves the following three questions:

- Does the paper focus on offensive security? If so, we assign it to the category "defense" (◐).
- Does the paper explore whether AI could be an effective hacking assistant? If so, we assign it to the category "assisted-hacking" ($\star$).

[6] We did not use MITRE ATLAS, since ATLAS focuses on vulnerabilities in AI-enabled systems, which are orthogonal to our focus (§II).

- Does the paper showcase a new attack? If so, we assign it to the category "attack" (†).

Third, we further explore two classes of papers.

If a paper proposed a novel attack (†), we examine whether the authors proposed any countermeasures ("Def."), resulting in a binary conclusion depending on whether they *explicitly* state such countermeasures. For instance, Anderson et al. [132] designed a deep learning-based DGA to intentionally bypass the DGA detector. Yet, they also explore approaches to enhance the detector's security.

If a paper is designed for offensive security (🛡), we check whether the authors *explicitly* stated that attackers could abuse the proposed technique. We report the offensive potential of a technique in the column "Pot. Abuse." One example of such an explicit statement is provided by Zennaro et al. [34] within the ethical considerations, acknowledging the potential for malicious use and condemning any such implementation to attack or harm.[7]

B. "What is the target of OAI?"

First, we evaluate whether the objective of the OAI use case is to impact a human (👤), a system (🗄), or both (👤+🗄).

- We categorize an attack as targeting a human (👤) if *(a)* it exploits a "vulnerability" in human behavior, such as susceptibility to misinformation (e.g., polarizing summaries [103]) or *(b)* it infringes on privacy (e.g., by inferring sensitive data from non-sensitive data [89]).
- If the attack is directed at a system (🗄), it implies that its success does not strictly depend on human involvement (e.g., bypassing an Intrusion Detection System [26], or CAPTCHA cracking [68].
- In instances where an attack is directed at both (👤+🗄) successful execution requires deceiving both human and system. For example, in a phishing scenario (e.g., [98]), success relies on the system failing to detect the phishing email and the human falling for the deception.

Furthermore, for attacks directed against a system (either 🗄 or 👤+🗄), we review whether the attack was directed against a real system (⚙) or a toy system (🤖). A toy system is a simplified, smaller-scale version of a real-world system used for experimentation. It can mirror specific aspects of a complex system with reduced user involvement and cost, as defined by Xu et al. [70]. For instance, Hu et al. [25] developed their own black-box detector, serving as a "toy system" to assess the proposed evasion technique.

Finally, to understand the role of the social aspect within each paper, we counted the occurrences of the words "society," "social," "societal," or "socio."

C. "What is the cost/benefit of OAI?"

First, we analyze each paper to assess *potential benefits* that may motivate attackers to utilize the proposed AI technique (benefit). Second, we evaluate each paper to identify any *associated costs* that could discourage attackers from adopting the proposed technique (cost). We scrutinize each paper for the below criteria (inspired by [62]):

(I) Did the authors analyze the *benefits* of the attacker leveraging the proposed AI technique? More precisely, did the authors provide any supplementary evidence/analysis/discussion of the attacker's benefits beyond simply stating that "the proposed method works"? Four answers are possible:
- No, no mention (🚫).
- Yes, qualitative (💬): Just a discussion. For instance, Iqbal et al. [79], in addition to showcasing technical use cases, discuss the benefits of ChatGPT for cybercriminals to refine their skills and facilitate more effective attacks.
- Yes, quantitative (🖩) (e.g., based on accuracy/precision). For instance, Ozturk et al. [81] evaluate static code analyzers and ChatGPT to detect OWASP vulnerabilities, comparing them based on success rate and accuracy.
- Yes, clear mention of monetary benefit or time/resources saved according to some metrics that go beyond sheer accuracy/precision (💰). E.g., Lee et al. [90] compare their technique with non-AI techniques to detect cross-site scripting vulnerabilities. They fixed the time budget to 3 hours, and then compare the number of detected vulnerabilities and the number of attack requests required to identify the vulnerabilities (required resources).

(II) Did the authors analyze the *costs* for designing/building/implementing the proposed technique? To examine this, we reviewed the papers individually and searched for keywords such as "cost," "money," "time," or "resources." Four answers are possible:
- No, no mention (🚫). E.g., Goldbeck et al. [23] leverage publicly available information to predict a user's personality, but do not mention the time/cost for data collection, algorithm development, runtime or similar.
- Yes, qualitative (💬): Just a discussion. As an example, Yu et al. [68] argue that their proposed method for CAPTCHA cracking is low-cost since: it uses an open-source library; it can be implemented on a personal computer; unlimited training is available.
- Yes, quantitative (🖩). This includes a clear numeric estimation of the costs based on some metrics, e.g., required resources or time (beyond sheer accuracy/precision). Liu et al. [113], for instance, compare wall clock time, and Pa Pa et al. [82] delineate that the development of functional malware with ChatGPT or AutoGPT, including debugging, takes around 90 minutes.
- Yes, clear mention of the required $$ to launch the attack (💰). We did not identify any paper in this category.

(III) For *Non-AI-baseline comparison*, we ask ourselves whether, e.g., the authors consider if the same objective could be achieved without AI. Three cases are possible:
- No, no mention (🚫) of any alternative.
- Yes, qualitative (💬). For instance, [108] provides a qualitative comparison to manual penetration tests.
- Yes, quantitative (🖩). For instance, Bahnsen et al. [127] quantitatively compare the effectiveness of DeepPhish in creating phishing URLs to non-AI techniques.

[7] The authors additionally refer to the Autonomous Weapons Open Letter: AI & Robotics Researchers of Future of Life.

Additionally, we check whether each paper shares the source code or prompts (📝) if an LLM was used. We include the link to the repository, if available.

D. Academic Literature (Techn. Papers): Additional Review

For the technical papers from the academic literature (discussed in §III-B) we additionally analyzed the *technical requirements* (e.g., algorithm, or data) to set up the attack. As identified by [62], most papers on adversarial machine learning do not consider the human effort required to technically implement the attack. Inspired by this observation, we extend the review of the technical papers by performing additional analyses, the results of which are presented in Table IV. Let us explain how we derived this table.

1) First, we scrutinize whether the applied algorithm/technique is publicly available (e.g., ChatGPT) and/or can be easily re-used (such as in [98]) If so, we mark the column with a "yes." Otherwise, if we do not mark the column, this means that the attacker has to develop the algorithm from scratch, such as in [127].[8]
2) If the algorithm needs to be developed from scratch, we review the availability of training data. We distinguish:
   - publicly available data (🖥), e.g., Biesner et al [84] use publicly available data sets of leaked passwords;
   - data collected by the authors (👤), e.g., Lee et al. [100] constructed six datasets of keyboard inputs;
   - and data collected by the authors with special access (👤*), e.g., Chung et al. [121] collected the operational data of building automation systems protecting a compute infrastructure—but they could only collect the data because they had *access* to such a system.

   Moreover, we examine whether shallow learning or deep learning techniques are used (also done in [62]). This allows a basic distinction between the required amount of training data and resources, since (deep) neural networks typically require more computational effort to be set up.

Finally, we review whether the authors publicly released their code and add the link to the repository, if available.

Appendix B
User Survey: Supplementary Information

We provide more low-level details on our user survey with non-experts (§V). We implemented the survey via Qualtrics for both web and mobile. Participation required user consent for anonymous data use; data was stored only after answering all questions and submitting the responses. Participants could exit and resume the survey at any time.

Demographics. The detailed demographics of the survey participants are illustrated in Figs. 6, 7, 8, and 9. We compared the demographics from our survey with the OECD demographics since 88% of the participants are from OECD countries. The goal is to identify any bias in our sample and include

TABLE IV: **Literature Review—Technical OAI Papers (Algorithm).** We report 79 technical OAI papers and the analysis of the *technical requirements* to set up the attack: We assess the availability of the applied algorithm (A-av.), —if the algorithm needs to be developed from scratch—the availability of training data (D-av.)), and whether deep learning techniques are used (DL). * indicates that special access for the data collection is required.

		Algorithm		
Paper	**Year**	**A-av.**	**D-av.**	**DL**
Antonelli [72]	2024		👤	✓
AlMajali [73]	2023		👤	
Chen [74]	2023		👤	✓
Chowdhary [75]	2023		👤	✓
Gallus [76]	2023	✓		✓
Ghanem [77]	2023		👤	
Happe [78]	2023	✓		✓
Iqbal [79]	2023	✓		✓
Karinshak [80]	2023	✓		✓
Ozturk [81]	2023	✓		✓
Pa Pa [82]	2023	✓		✓
Zennaro [34]	2023		🖥	
Auricchio [83]	2022		🖥	✓
Biesner [84]	2022		🖥	✓
Cody [85]	2022		👤	✓
Confido [86]	2022		🖥	✓
Gangupantulu [87]	2022		🖥	✓
Hu [25]	2023		👤	✓
Jagamogan [88]	2022	✓		
Karanatsiou [89]	2022		👤	
Lee [90]	2022		👤*	✓
Li [91]	2022		👤	✓
Lin [26]	2022		🖥	✓
Nhu [92]	2022		👤	✓
Pagnotta [93]	2022		🖥	✓
Tran [94]	2022		👤	✓
Yao [95]	2022		👤	✓
Caturano [96]	2021		👤	
Erdődi [33]	2021		👤	✓
Gangupantulu [97]	2021		🖥	✓
Khan [98]	2021	✓		✓
Kujanpää [99]	2021		👤	✓
Lee [100]	2021		👤	
Maeda [101]	2021		🖥	✓
Neal [102]	2021		👤*	
Sharevski [103]	2021		🖥	
Standen [104]	2021		🖥	✓
Toemmel [105]	2021		🖥	✓
Tran [106]	2021		🖥	✓
Al-Hababi [107]	2020		👤	
Bhattacharya [37]	2020		👤	
Chowdhary [108]	2020		👤	✓
Halimi [109]	2020		👤	
Hu [110]	2020		👤	✓
Lee [111]	2020		👤	
Lee [112]	2020		👤	
Liu [113]	2020		🖥	✓
Pearce [114]	2020		👤	
Sharevski [115]	2020		👤*	
Shu [116]	2020		🖥	✓
Song [117]	2020		👤	
Valea [35]	2020		🖥	
Yu [118]	2020		🖥	✓
Basu [119]	2019		👤*	✓
Cecconello [120]	2019		👤	
Chung [121]	2019		👤*	
Das [122]	2019		👤	✓
Ghanem [123]	2019		👤	
Tshimula [124]	2019		👤*	
Yu [68]	2019		🖥	✓
Zhang [125]	2019		👤	✓
Anand [126]	2018		👤	
Bahnsen [127]	2018		🖥	✓
Kronjee [128]	2018		🖥	
Rigaki [129]	2018		👤	✓
Zhuo [130]	2018		🖥	✓
Yao [131]	2017		🖥	✓
Anderson [132]	2016		🖥	✓
Ceccato [36]	2016		👤	
Grieco [133]	2016		🖥	
Freitas [134]	2015	✓		
Bursztein [135]	2014		👤	
Adali [136]	2012		👤	
Malhotra [137]	2012		👤	
Sumner [138]	2012		👤	
Goldbeck [23]	2011		👤	
Yamaguchi [139]	2011		🖥	
Bursztein [140]	2009		👤	
Golle [49]	2008		👤	✓

[8] Khan et al. [98] re-use a pre-trained large language model to generate spear-phishing emails, while Bahnsen et al. [127] design DeepPhish, a Long Short-Term Memory Network, that learns the structure of effective URLs and then generates new synthetic URLs to create "better phishing attacks.“

potential limitations when interpreting the results. We focus on the OECD since this is the largest group of countries covering our participants. The remaining 12% are from 17 distinct countries – with one participant each from 14 different countries, 21 from India, 20 from Liechtenstein, and 4 from Qatar.[9] The comparison of the participants' demographics from OECD countries to the overall population from the OECD is reported in Table V.

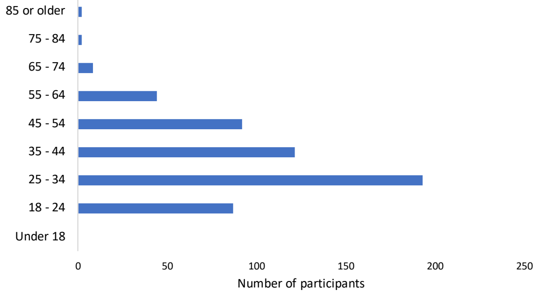


Fig. 6: **Age range.** 51% of the participants are below 35, and overall 90% of the participants are between 18 and 54. Only 2% are older than 65.

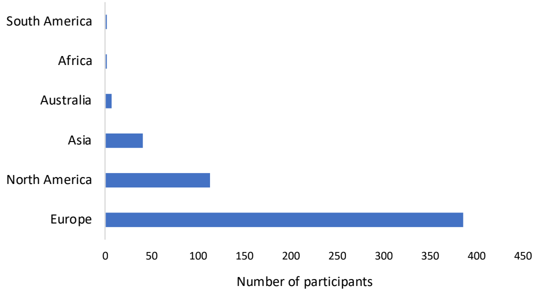


Fig. 7: **Residence based on continent.** Most participants are from Europe (70%) and North America (21%). 7% are from Asia, while overall only 2% of all participants are from Australia, South America, or Africa.

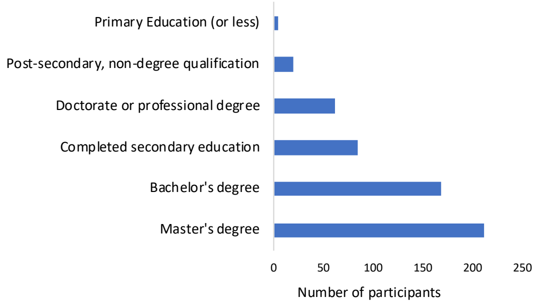


Fig. 8: **Education.** 20% of the participants do not hold a university degree, while 31% hold a Bachelor's degree, and 50% a Master's degree or higher.

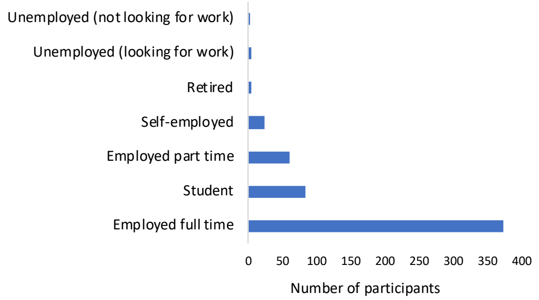


Fig. 9: **Employment.** The majority of the participants are employed (68%). 15% are students and 10% are self-employed. 2% are retired or not employed.

TABLE V: **Comparison of our sample to OECD population.** The data is reported in %. Compared to the OECD reference values, our sample includes fewer women and fewer individuals over 55. Also, our sample includes a comparatively high number of highly educated participants.

	Survey Part. from OECD	**OECD Reference**
Gender		
Female	36.02	50.77
Male	63.15	49.23
Other	0.83	N/A
Age		
18 - 24	15.32	11.13
25 - 34	34.58	16.83
35 - 44	22.77	16.95
45 - 54	16.56	16.69
55 - 64	8.70	15.69
65 - 74	1.66	12.38
75 - 84	0.41	7.32
85 and over	N/A	3.02
Education		
Primary	0.75	20.00
Secondary	11.03	40.00
Tertiary	88.22	40.00

Expertise. Figs. 10 and 11 provide details on the participants' expertise in AI and cybersecurity. We asked the users to rate their knowledge of each field as either of the following:

a) Beginner ("I have little or no knowledge of AI/cybersecurity.")
b) Intermediate ("I have a basic understanding of AI/cybersecurity concepts.")
c) Advanced ("I have a solid understanding of AI/cybersecurity and its applications.")
d) Expert ("I have extensive knowledge and experience in AI/cybersecurity.")

We asked participants for a self-assessment of their knowledge to put their responses into further context based on their background. We chose not to assess participants' knowledge as it would add complexity to the survey and is not essential for our primary objective, i.e., collecting public opinions on the malicious use of AI. Indeed, prolonging the survey completion time rather contradicts our objective.

Completion Statistics. Qualtrics recorded 596 survey initiations, with 82% completions. Dropouts mainly occurred either directly after giving consent or just before the requirement to provide qualitative input. On average users took four minutes to complete the survey (excluding outliers[10]). The survey was open for four months; however, most participants provided their answers in September 2023 (68%), some in October (7%), November (14%), and December (11%).

Correlation Analysis. For the correlation analysis, we consider the user demographics as independent variables and the provided responses to the main body of the survey as

[9] We rely on the data from the OECD demographics database as of 2022 – the latest available data. We refrain from analyzing the demographics from India, Liechtenstein, and Qatar due to the limited sample size.

[10] Users taking more than 20 minutes are outliers: they likely did not complete the survey in a single session but instead paused and resumed.

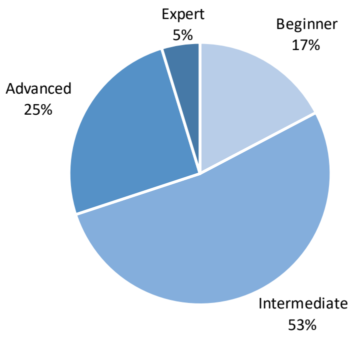


Fig. 10: **AI Expertise.** 17% of the participants have little or no knowledge of AI, while around half of the participants have a basic understanding of AI, and only very few are experts.

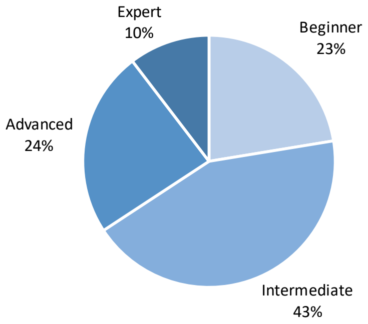


Fig. 11: **Cybersecurity Expertise.** Only 23% of the participants have little or no knowledge of cybersecurity, while the majority of participants have at least an intermediate understanding of cybersecurity.

dependent variables. First, we transform the qualitative responses into quantitative ones (i.e., we map "yes" or "no" to "1" or "0"; whereas the knowledge level of AI/cybersecurity is represented by ordinal numbers, e.g., "beginner" to "1" and "intermediate" to "2"). Second, we perform the Pearson correlation method, one of the most common ones [222], and derive the corresponding statistic (in §V-B1).

How is it possible that some respondents who are not concerned about OAI think that AI will harm them? In Fig. 5, it is intriguing to observe that there are 83 people who are "not concerned about OAI" and, among these, 21 believe that "AI will harm me." There is a plausible explanation for this apparently nonsensical outcome. Indeed, these people do not worry about AI being used offensively—in the sense that they are not afraid (or are unaware) of the fact that cybercriminals may use AI to cause harm to them; however, these people may still be worried that AI may harm them because of, e.g., unintended AI failures or job displacements due to AI. As a matter of fact, our qualitative analysis found that many respondents provided similar concerns—which have little to do with cybercriminals. This is evidence that the term "offensive AI" may not be fully understood by some individuals, or they simply have more pressing concerns about AI which may cloud their vision about OAI-related problems.

Appendix C
User and Expert Survey: Qualitative Analysis

We provide supplementary information on the qualitative analysis of the user survey (in §V) and the expert survey (§VI-A) based on the constructivist grounded theory [57] enhanced through two rounds of axial coding—termed horizontal and vertical coding. This methodology incorporates principles of constructivism and pluralism, leading to a more inclusive and comprehensive data analysis.

Codebook. We developed an ad-hoc codebook for our qualitative analyses (§II-C), which we derived after inspecting the responses to the (mutually exclusive) questions "what are you most concerned about OAI?" and "why are you not concerned about OAI?" of our survey. We present our codebooks in Table VI (Table VII), showing the horizontal axial codes (perspectives), along with their corresponding vertical axial codes and the associated related focused codes for each axial code for people who are (or not) concerned about AI.

Application. After deriving our codebook, we then apply it to analyze the corresponding responses—for both the non-expert survey (§V-B2), as well as for the expert survey (§VI-A). We report the results in Table VIII (Table IX), presenting the focused codes derived from the qualitative analysis, representing people's reasons and dimensions for being concerned about AI (not being concerned about AI). These focused codes were generated on the basis of the initial codes, with the frequency indicating how often each focused code emerged in relation to the initial codes.

Appendix D
Expert Statements: Source and Analysis

In Appendix D-A, we list the statements provided by the experts, as indicated in our methodology (discussed in §II-D). Then, in Appendix D-B we provide low-level details explaining how we objectively analyzed them (refer to §VI-B1).

A. Statements (Verbatim)

We report the statements as written by each expert. The order is random, and we will not provide any information that can be used to identify the author of each statement.

Statement from Expert 1: *Is it AI-generated content or reality?* To me, the scariest aspect of offensive AI is related to deepfakes and disinformation. Nowadays, we *cannot trust* anything we see, hear, or read online anymore. As soon as I see a picture or video, I have to wonder: "is it real or was it AI-generated?" It is becoming increasingly harder to distinguish AI from reality. This forces individual users to take time to, e.g., look at a picture or video, and decide if it is real or not. Unfortunately, Web users do not always challenge what they encounter online: if they have seen a picture showing a specific fact, then they may assume that this fact definitely took place.

Attackers now use AI to create disinformation campaigns, fake reviews, and to manipulate crowds into believing that some AI-generated content is real. This has direct political and economical implications, e.g., election campaigns. One of the

TABLE VI: **Axial Coding—Concerned.** Horizontal axial codes, vertical axial codes and related focused codes for people concerned about AI.

Horizontal Axial Codes	Vertical Axial Code	Description	Related Focused Codes
AI Impact and Societal Concerns	Ethical Implications of AI	Concerns about the ethical challenges posed by AI.	AI in Healthcare, AI Surpassing Human Performance, Ethical Concerns, Military Applications and Warfare, Unintended Errors
	AI's Societal Impact	Apprehensions about AI's broader societal effects.	Negligent AI Development Practices, Loss of Control in Autonomous AI, Cybersecurity Attacks, Reduction in Human Learning, General Concerns, Job Displacement, Regulatory Deficiencies, Military Applications and Warfare, Spread of Misinformation, Adverse Social Impacts, Speed, Automation, and Ease of Use, Unintended Errors
	Privacy and Identity Concerns	Fears related to privacy, identity theft, and digital manipulation.	Privacy, Profiling, and Manipulation, Data Misuse, Identity Theft and Deepfakes
Technological Risks and Threats	Risks of Advanced AI Capabilities	Concerns about AI surpassing human capabilities and control.	AI Surpassing Human Performance, Loss of Control in Autonomous AI, Reduction in Human Learning, Job Displacement, Speed, Automation, and Ease of Use
	Cybersecurity and Data Risks	Fears of AI being exploited for cyberattacks or data abuse.	Privacy, Profiling, and Manipulation, Data Misuse, Cybersecurity Attacks, Identity Theft and Deepfakes
	AI Development and Error Risks	Concerns regarding errors and issues in AI development practices.	AI in Healthcare, Negligent AI Development Practices, Unintended Errors
AI in Domain-Specific Contexts	AI in Healthcare	Concerns about AI's application in healthcare settings.	AI in Healthcare
	AI in Military and Security	Apprehensions regarding AI's use in military and warfare contexts.	Military Applications and Warfare
	AI's Impact on Work and Skills	How AI affects job security and human capabilities.	AI Surpassing Human Performance, Reduction in Human Learning, Job Displacement, Speed, Automation, and Ease of Use
Human-AI Interrelations	AI Autonomy and Supremacy	Concerns about AI surpassing human capabilities and going out of control.	AI Surpassing Human Performance, Loss of Control in Autonomous AI, Speed, Automation, and Ease of Use
	Job Displacement by AI	Concerns about AI replacing human jobs.	AI Surpassing Human Performance, Job Displacement, Speed, Automation, and Ease of Use
	AI's Influence on Humans	The impact of AI on human learning and cognitive abilities.	Reduction in Human Learning, General Concerns
AI Development and Regulatory Issues	AI Development Practices	Issues and concerns in the processes of developing AI.	Negligent AI Development Practices
	Regulatory Frameworks for AI	The need for and absence of adequate AI regulation.	Regulatory Deficiencies, General Concerns
	Operational Challenges of AI	Practical challenges in the deployment and operation of AI.	Data Misuse, Speed, Automation, and Ease of Use, Unintended Errors, Loss of Control in Autonomous AI

TABLE VII: **Axial Coding—Not Concerned.** Horizontal axial codes, vertical axial codes and related focused codes for people not concerned about AI.

Horizontal Axial Codes	Vertical Axial Code	Description	Related Focused Codes
Technological Progression and Safety Perspective	Evaluation of AI Maturity	Assesses the current development stage and future potential of AI.	AI Not Yet Mature
	Safety and Risk Management in AI	Focuses on the safety concerns and risk management strategies in AI.	Adequate Protection and Prevention
	Balancing Progress and Safety	Examines the balance between technological progression and safety.	Balanced Perspectives, Predominance of Benefits
Societal and Ethical Implications Perspective	AI's Societal Impact	Explores how AI impacts society, including its benefits and harms.	Balanced Perspectives, General Lack of Concern, No Novel Concerns
	Ethical Considerations in AI	Discusses the ethical concerns related to AI development and use.	Concerns Centered on Humans, Not AI, Adequate Protection and Prevention
	Public Attitudes towards AI	Analyzes public sentiment and attitudes towards AI.	AI Not Yet Mature, Predominance of Benefits
Perceptual and Comparative Analysis Perspective	Comparative Analysis of AI	Weighs AI's benefits against its potential harms.	Balanced Perspectives, Predominance of Benefits
	Public Nonchalance about AI	Captures the general lack of concern about AI's impact.	General Lack of Concern, No Novel Concerns
	AI Readiness and Perception	Assesses perceptions of AI's readiness and potential benefits.	AI Not Yet Mature, Predominance of Benefits
Human-Centric Approach Perspective	Human-Centric AI Impact	Focuses on AI's impact from a human-centered perspective.	Concerns Centered on Humans, Not AI, Predominance of Benefits
	Perception of Protection Measures	Examines beliefs regarding the adequacy of AI protection measures.	Adequate Protection and Prevention
	Balancing Human Concerns and AI Advancement	Weighs human concerns against the perceived advancements in AI.	Balanced Perspectives, Predominance of Benefits

TABLE VIII: **Focused Codes—Concerned.** We present the focused codes and their frequencies w.r.t. the initial codes for people concerned about OAI.

Focused Code	Frequency	
	User Survey	Expert Survey
AI in Healthcare	4	0
AI Surpassing Human Performance	14	0
Privacy, Profiling, and Manipulation	19	3
Negligent AI Development Practices	8	0
Data Misuse	12	0
Loss of Control in Autonomous AI	20	0
Cybersecurity Attacks	141	4
Reduction in Human Learning	15	0
Ethical Concerns	19	0
General Concerns	24	0
Identity Theft and Deepfakes	61	1
Job Displacement	16	0
Regulatory Deficiencies	9	0
Military Applications and Warfare	21	0
Spread of Misinformation	98	0
Adverse Social Impacts	11	0
Speed, Automation, and Ease of Use	11	1
Unintended Errors	10	0

TABLE IX: **Focused Codes—Not Concerned.** We present the focused codes and their frequencies w.r.t. the initial codes for people not concerned about OAI. (No participant of the expert survey was not concerned about OAI.)

Focused Code	Frequency User Survey
AI Not Yet Mature	5
Adequate Protection and Prevention	21
Balanced Perspectives	11
Concerns Centered on Humans, Not AI	9
General Lack of Concern	15
Predominance of Benefits	8
No Novel Concerns	11

most important open problems in the field of OAI is therefore to be able to detect such deepfakes and disinformation campaigns.

Adverse effects of AI. Second, AI was initially meant to *make our life easier*. Unfortunately, it also makes an attacker's job easier, for example by using AI to write sophisticated phishing emails or advanced malware. A few years ago, phishing emails were quite easy to detect, e.g., because they were poorly written, full of grammar mistakes, and impersonal. Now, with ChatGPT and other LLMs, they look more authentic. Given that Web users also use such services to write (benign!) emails faster, it is becoming more challenging to detect such phishing emails and other AI-generated threats. While AI was meant to simplify some day-to-day tasks, it is also introducing new threats and challenges. A second open problem revolves around weighting the pros and cons: *can an AI-based solution be misused?* It is critical to take into account potential adverse effects of AI before deployment, otherwise we end up in a world that is harder and more dangerous to live in than it was before AI.

Privacy-related attacks. Third, AI can also be misused to collect or infer sensitive information about Web users. For example, AI facilitates profile-matching across social networks, enables attribute-inference attacks, and deanonymization online. It can also have direct military implications when AI is leveraged for (unauthorized) surveillance. A third open problem consists in developing guidelines and defensive measures to better protect users and their (sensitive) data, but also to work on awareness campaigns to let users know about AI-based privacy-related threats.

Statement from Expert 2: Given the rapidly increasing prevalence of AI, especially generative AI, in all walks of life, offensive AI is becoming an increasingly important topic. Within existing research, a rather positive image of AI often prevails. Discussions tend to focus on the potential positive aspects and impacts of AI. Although the dangers of AI are more and more being discussed, such discussions are often limited to side effects arising from predominantly benign use of AI, such as biases in benign AI systems or privacy violations during AI model training for benign AI systems. The active use of AI for malicious purposes has so far played a minor role in academic discourse but deserves more of our attention. In particular, I see the following three open problems as central challenges of offensive AI research that the academic community should address more closely.

First, research should specifically focus on offensive AI in the field of high-risk AI applications. One can certainly argue that offensive AI should generally be classified as a high-risk AI application. However, I believe that there are particularly high-risk areas of application for AI where offensive AI can be especially harmful. This applies in particular to applications of offensive AI that jeopardize democratic institutions and institutions, which uphold the rule of law (e.g. spreading false information to manipulate elections with the help of AI), or critical infrastructures (e.g. AI-based ransomware in medical facilities).

Second, from a socio-technical perspective, research should focus more on the interaction between humans and technology in terms of offensive AI. For example, research should delve deeper into how individuals and organizations can detect and counteract offensive AI attacks. This includes investigating what makes individuals and organizations particularly vulnerable to offensive AI and developing tools to help them recognize and fend off attacks with offensive AI (e.g., email clients that detect texts generated by generative AI and alert users).

Third, research should explore how fundamentally benign AI can turn into offensive AI and be used for malicious purposes. Typical examples include generating high-quality scam emails using generative AI. It is not surprising that fundamentally benign AI can also be misused for malicious purposes. Therefore, we should engage more in the systematic study of the potential dangers of benign AI in terms of its use as offensive AI and develop appropriate tools for assessing the risk potential for AI developers and researchers.

Statement from Expert 3: In my opinion the primary open problems in the field of Offensive AI at this point in time are:

1) Lack of understanding regarding the potential and limitations of Offensive AI tools. One of the greatest problems in this field is the lack of understanding regarding the real potential and limitations of AI on the offensive side. The vast majority of the published research presents success in achieving a specific goal (e.g., to jack a webserver using an AI tool) which stimulates our thinking regarding the potential of Offensive AI. On the other hand, research that shows failures and sheds light on the limitations of Offensive AI tools has yet to be published. A greater understanding of the boundaries of Offensive AI tools is required to shed light on the real potential and limitations of Offensive AI tools concerning integrated AI technology.

2) Lack of Effective Countermeasures. In some usecases, AI has reached the maturity required to orchestrate cyber attacks that could lead to financial losses (e.g., audio deep fakes). However, despite the rapid advancement made by the community on the offensive side, one of the greatest problems in this field is the fact that there needs to be more advancement on the opposite side of countermeasures. There are usecases in which effective countermeasures that could be used to detect, mitigate, and prevent AI-orchestrated attacks were not developed (e.g., detecting fake news). In other cases, the countermeasures that were developed were found ineffective in reality when simulated in real environments (e.g., audio deepfakes). There is a real need to develop countermeasures that are effective against the current threats posed by Offensive AI in real environments (e.g., audio deepfakes, text deepfakes, etc).

3) Lack of effective TARA standards for systems against Offensive AI. One of the greatest problems in this field is the inability to determine whether the risk posed to systems by the advancements published by new research is real. This happens because the practicality and the real outcome of the attacks are not clear from the research published in this field due to the facts that: (i) the vast majority of the research hasn't been demonstrated against real systems and it is known that attacks against specific components may not affect the system itself, (ii) the level of expertise required to use the tools is not clear (layman? expert?). This requires the development of dedicated threat analysis and risk assessment standards (TARA) that will allow CISOs to objectively calculate the risk posed by offensive AI tools to their systems, taking into account the Technological Readiness Level of a tool, the practicality of the attack (considering expertise, needed equipment, the window of opportunity, etc.), and the outcome of the attack in practice (considering validation against real systems). Such standards may help to quantify and shed light on the real risk posed by the developed offensive tools to organizations, putting things into the right perspective.

Statement from Expert 4: After reading the draft version of the SoK paper, I strengthened my conviction that humans are the most relevant target of offensive AI and, therefore, future research on offensive AI should prioritize the offensive use of AI to hack human beings by exploiting the large amount of data that can be collected on human behaviors and habits.

Therefore, in the following, I briefly describe the open problems in the field of offensive AI that I would like to give priority:

1) Cognitive biases and offensive AI. We have already clear evidence that AI can be used to hack human by performing contextualization and personalization in phishing attacks. It is therefore easy to conjecture that offensive AI could exploit cognitive biases of individuals to accomplish tasks that violate security and privacy. As an example, confirmation bias is a well-known and powerful cognitive bias that affects decision-making. Offensive AI could exploit this bias by creating echo chambers where users are only exposed to information that confirms their existing beliefs. This can lead to increased polarization, reduced critical thinking, and the spread of misinformation.

2) Offensive AI and behavioral economics. The field of behavioral economics pointed out well how cognitive biases strongly affect the financial decisions of individuals and how these decisions deviate from those predicted by classical economic theory. Given the above conjecture that offensive AI can exploit cognitive biases of individuals, it is a priority to investigate how offensive AI could be used to manipulate financial decisions of individuals at a large scale with serious financial consequences for the society. As an example, loss aversion is a well-known cognitive bias where the pain of losing is felt more acutely than the pleasure of gaining. Offensive AI could exploit this bias by spreading information targeted to each individual that emphasize potential losses, so prompting individuals to make irrational financial decisions to avoid these perceived losses.

3) Human-AI Teaming against Offensive AI exploiting cognitive biases. I do believe that as offensive AI will become more sophisticated in exploiting cognitive biases such as confirmation bias and loss aversion, there will be a pressing need for defence measures that leverage both AI and human strengths. In fact, human-AI teaming can combine the rapid processing and pattern recognition capabilities of AI with contextual understanding and decision-making of humans. This line of research could intersect with the previous one on offensive AI and behavioral economics, given the seminal work of the Nobel Laureate Daniel Kahneman in behavioral economics on the analysis of comparative strengths and weaknesses of humans and machines in decision making; this previous work could be leveraged to combine humans and AI capabilities against offensive AI exploiting cognitive biases.

Statement from Expert 5: Offensive AI can be used to cause harm to individuals, institutions and society at large. Among the open problems in the field of Offensive AI, the top three which need to be researched and solved are: detection of deepfakes, mitigating the use of AI to spread disinformation, and developing defensive AI mechanisms to counter AI-orchestrated cyberattacks.

Deepfake technology enables attackers to launch phishing and other forms of social engineering attacks by impersonating a victim through cloning their face or their voice. The variety of ways by which this can cause threats and harm makes it one of the most serious threats of Offensive AI. Deepfakes can be used to overcome biometric systems. They can be used to gain access to information and data. Attackers can use the technology to carry out fraud, to launch disinformation campaigns that seem credible, or for defamation. Mechanisms need to be developed to detect deepfakes to counteract their threat.

AI combined with social media can be used to accelerate the spread of disinformation for a variety of ulterior motives, such swaying public opinion and shaping behavior. Renee DiResta, of the Stanford Internet Observatory, says in relation to disinformation: "social media took the cost of distribution to zero, and generative AI takes the cost of generation to zero" [223]. New technology needs to be developed and incorporated into social media platforms in order to detect and prevent the spread of disinformation. This will not happen organically. It will only be propelled by the force of government legislation.

Offensive AI can be used to enable new levels of cyberattack automation from launch to the various stages of attack propagation. It allows attackers to scale up their attack coverage and increase their success rate. This renders human controlled detection systems incapable of keeping up with the scale and speed of the attacks. Defensive AI mechanisms need to be developed which are capable of automatically detecting and counteracting their malicious twins.

Statement from Expert 6: Open problems with offensive AI can be categorized into areas of focus based on the motivations of different expert groups. These groups may include offensive teams aiming to improve attacks, defense teams aiming to mitigate attacks, and trust and safety teams aiming to minimize harm to society. For red team purposes, the challenge lies in measuring the success and value of offensive AI attacks, as it may not always offer substantial increases in efficacy over traditional techniques. For blue teams, the challenge may be that offensive AI simply enables relatively trivial attacks to be scaled, with offensive AI being used to turn large groups of lower skilled adversaries into more highly skilled ones. For civil society, the challenge will be keeping up with new approaches, such as information/influence operations, and addressing the societal challenges they pose.

For offensive teams themselves, an obvious challenge is how to measure the success and value of attacks powered by offensive AI. It is not simply a case that offensive AI will always offer substantive increases in efficacy over other more traditional offensive techniques and if this is not the case, then questions should be raised as to the value of the approach. In particular, just as with other quantitative and qualitative measures of success, the value of AI can often be manipulated to show success and/or failure as desired. An example of this is that there is already research that purports to show how generative algorithms can find ”0-days”. This research needs to be reviewed against results using traditional non-AI based techniques, to determine whether the approach delivers the value initially claimed. The challenge can be summarised as finding suitable cross-domain experience and mechanisms to allow effective comparison and scoring of traditional and AI-powered offensive techniques.

For those trying to defend against existing adversaries, the challenge may simply be that offensive AI will likely be used to enable scaling of attacks. When we look at DoS for example, there are both technical attacks but also things that simply overwhelm the human at the other end of the process. What happens if it turns out that offensive AI is most effective when used to turn large groups of lower skilled adversaries into more highly skilled adversaries? For example, what tools and techniques will defenders need to spot and deal with these attacks? Assuming that generative algorithms are successful at complying with the prompts provided by their human users then their output will likely pass the requirements enforced by traditional security controls. This challenge can be summarised as finding ways to classify how particular offensive AI techniques may be weaponised and then defining solutions that are appropriate to each.

For society at large, the challenge will be less about measuring and improving efficacy of offensive AI usage or stopping individual attacks but rather trying to keep up with new approaches. Offensive AI will likely result in new types of techniques that don’t fit conventional cyber security definitions. For example, are information/influence operations considered as part of offensive AI and how can they best be addressed?

Statement from Expert 7: Researchers are beginning to develop AI for various offensive use cases that will present challenges to defensive systems and processes in the potency and speed of cyber campaigns via offensive copilots, scaling social engineering attacks, and enhancing offensive operations. This statement outlines these challenges and the gaps that need to be addressed for these techniques to be considered effective.

Offensive Copilots: Reducing Time to Impact. Offensive copilots are AI systems that can significantly reduce the time required for threat actors to execute an attack by automating labor-intensive tasks that typically demand specific expertise or extensive exploration. For example, AI can expedite the discovery, development, and delivery of exploits or polymorphic malware. The primary gap for attackers in achieving this capability is largely an engineering problem of agents systems to tools (e.g., reversing tools, CVE lookups, executable instrumentation) through agent systems, checks to validate successful outcomes, and the tedious optimization of system instructions in generative AI for the copilots actions to produce reasonable results.

Scaling Social Engineering Attacks with Minimal Direction. AI’s potential to scale social engineering attacks is another developing attack vector. Threat actors can use generative AI to create deepfakes and other convincing forms of fake identities, which can be used in automated and interactive phishing or scamming operations. Unlike other automated social engineering tools, the AI-driven attacks are adaptive and interactive. The gap for attackers in achieving this lies in the engineering of agent-based systems with optimized playbooks for carrying out such operations. To be convincing, especially in interactive situations, further research is needed in realistic interaction techniques. This includes developing scoped questions and responses, achieving alignment without providing the idiosyncratic generic responses of AI like ChatGPT, and mimicking human-like selectivity in responses, including delays in responding. The primary research required is to enhance these systems’ realism and effectiveness in social engineering contexts.

Scaling Offensive Operations. Offensive AI enhances the targeting and scaling of cyber operations beyond traditional automation. AI-driven attacks can achieve higher precision and impact by integrating multiple automated steps from frameworks like MITRE ATT&CK, which maps out various cyber adversaries’ tactics and techniques. By integrating generative AI agent frameworks with preexisting tools, attackers can orchestrate complex operations that tackle large portions of an attack lifecycle seamlessly. However, achieving the scale and sophistication required for these operations presents significant engineering challenges. Additionally, there may be insufficient data on complex attacks to generate novel end-to-end attacks effectively. Fine-tuning large language models (LLMs) to cyber-operations might be necessary so that the AI can selectively focus on key indicators in long contexts and predict appropriate next steps accurately.

Statement from Expert 8: In the realm of threat intelligence research, state-sponsored threat actors have been reported to leverage offensive AI in their attack campaigns[11]. These actors use AI services for reconnaissance (identifying targets, researching tools, translating technical papers) and weaponization (scripting, creating social engineering content).

From this perspective, several open research questions remain. One question is whether offensive AI has been employed in other stages of the cyber attack kill chain [224]. For example, while there has been research on using AI for code obfuscation, it is unclear if threat actors have adopted AI-obfuscated malware in the wild. Can threat actors enhance their fuzzing capabilities with neural networks to find software vulnerabilities? Defensive Endpoint Detection and Response (EDR) products use AI to analyse logs and construct attack paths and timelines—can AI also be leveraged to reverse this process? Can threat actors use AI to design and automate lateral movement within compromised networks? These are intriguing questions that warrant further exploration.

Another important question is whether we can detect and mitigate attacks launched with offensive AI tools, and how these approaches differ from existing detection techniques. For instance, can AI-model obfuscated malware still be detected through entropy measurements[12]? Can phishing emails generated by AI be identified through content analysis?

Lastly, there are critical questions regarding AI service providers. What techniques can detect and prevent platform abuse? To date, there have been few reports from AI service providers about their tools being used for offensive attacks. Despite potential conflicts of interest with company reputation, increased transparency and disclosure would benefit the industry and the public, leading to stronger collective defense. Beyond detection techniques, what forms of collaboration or communication channels should be established between service providers and industry partners? What information can be shared to enhance collaboration? These topics are essential for ongoing discussion. In conclusion, while offensive AI might present more challenges for the defenders, it also opens up new avenues for research and collaboration. Addressing these open questions will require a concerted effort from researchers, industry professionals, and AI service providers.

Statement from Expert 9: Technological advancements are usually considered in terms of how they can be used by humans to benefit their tasks or life in general. This common approach towards understanding the positive effects of technology has in particular dominated inquiries in the interdisciplinary field of Information Systems (IS), which traditionally investigates the use of digital systems from a socio-technical perspective. Only in recent years this attention has somewhat shifted towards investigating the "dark side" of technology to understand adverse usage outcomes, albeit still mainly considering the (non-)achievement of predetermined goals. There is an emerging stream of research in IS on responsible AI with a particular emphasis on principles needed to mitigate AI related risks. Notwithstanding, the current discourse in IS has largely overlooked the offensive potential of AI, as considered in this study, which demonstrates that AI clearly provides or improves capabilities of adversaries for many unknown or weakly understood use cases that have the potential to disrupt not only the lives of people but also organizations and entire societies.

The following three research problems represent broad OAI capabilities that can be seen most threatening from a cost/benefit perspective, as they offer high rewards (or harm) and little cost to adversaries in terms of enabling or enhancing their attacks, and are difficult to mitigate from the defenders' standpoint. These include (1) social engineering, (2) information gathering and misinformation, and (3) vulnerability exploitation. Regarding (1), humans have been generally seen to be the weakest link in private and organizational contexts. AI will most likely aggravate this situation by, e.g., allowing for quickly building fake personas/profiles that can be used in spear phishing or other attacks on targets that have been cost-effectively identified via AI. From the IS perspective, open research issues comprise, e.g., how humans process or act on these attacks (e.g., heuristically vs. systematically), safeguards or countermeasures such as awareness building and associated training methods, or the use of automated AI systems to counter AI-enabled attacks. Regarding (2), information gathering and dissemination also scale well via AI and can provide the basis for a number of attacks to, e.g., use AI for reconnaissance, which is particularly hard to prevent, and efficiently disseminating content crafted or fabricated by AI that can be used to, e.g., polarize or reduce trust in institutions or science, thereby adversely affecting public opinion. Regarding (3), from an adversarial perspective an attacker can abuse AI systems by exploiting AI vulnerabilities through white-box, black-box, or gray-box techniques, which at least partially can be achieved with little technical knowledge. From the IS perspective, important safeguards to investigate may for example include data governance practices to prevent unauthorized access to sensitive (training) data and data biasing.

These research problems are of course neither exclusive nor exhaustive. From a more holistic view, these and other issues necessitate research on the effective government of AI, accounting for all stakeholders (as shown in this study), which should help mitigating many problems. For example, it will be necessary to investigate the role of global laws and regulations, and human oversight in dictating and testing the functions of AI, which in turn should reduce the capabilities of AI for offensive purposes in many use-cases. Also, respecting fairness, transparency and accountability (FAT) principles, currently debated in work on responsible AI, in the early design stage of AI systems development, e.g., through guiding frameworks or standards would constitute another important safeguard. In any case, a cautious and step-wise approach toward the implementation of AI-based solutions in organizations and society, providing for sufficient time for testing and corrections, is highly advised considering the potential of

[11] https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

[12] https://redcanary.com/blog/threat-detection/threat-hunting-entropy/

OAI.

Statement from Expert 10: In my opinion, these are the three most important problems:

1) Measuring AI usage in real-world attacks. An important open question is how to reliably measure the use of offensive AI in real-world attacks. So far, the community has studied a range of theoretically possible AI attacks, but we don't have sufficient understanding of whether and how often such attacks are carried out by attackers in practice, who are running these attacks, and what strategies attackers have to integrate/use AI. These questions cannot be answered without conducting empirical measurements by collecting and analyzing real-world attack data. Answering these questions is critical to the community to define accurate threat models and develop effective countermeasures. A key technical challenge is to distinguish AI-generated offensive content (text, images, voices, network traces, software code) within real-world data with limited "ground truth". Whether it is malware samples, network traces, or data from social media, it may be possible to manually label "offensive content" (e.g., malware) but it's still challenging to further label the AI-generated offensive content (e.g., AI-written malware) from offensive content generated by traditional methods (e.g., malware written by hackers). Solving this problem is key to running real-world measurements to plot out the threat landscape.

2) Defending against AI use in disinformation and online deception. Disinformation is a major threat to our society today, and it is difficult to address this threat with technical means only. Disinformation has been a problem even before the take-off of generative AI (e.g., with manually crafted fake news, and manually altered photographs). The key difference is generative AI significantly reduces the technical barriers/costs of generating false content, which, to some extent, democratizes cyber offense. Today, lay users (without any programming experience) are able to craft high-quality "deepfakes" using basic text prompts. In addition to detecting such AI-generated content, a bigger challenge is to determine the "intent" of the deepfake, effectively flagging the content with a malicious intention from that of benign-intended (e.g., AI content for entertainment). This process may require collaborative efforts from human users/moderators and automated detection techniques.

3) Using offensive AI to enhance existing defense. Offensive AI has the potential to be used positively to improve our defense. A concrete example is to use AI methods (e.g., Large Language Models or "LLM") to scan software code bases to detect bugs/vulnerabilities and augment traditional tools such as fuzzing. However, as an offensive technique (for vulnerability discovery), it can also be used by malicious parties to find zero-day vulnerabilities to facilitate system compromise. The question is how to tip the scales to benefit the defenders even more.

Statement from Expert 11: *1) Privacy attacks beyond membership inference.* One of the most well-researched uses of AI to attack privacy are attacks on training data, including membership and attribute inference [225]. However, AI can also enable other privacy attacks, such as linking of separate data items, e.g., for cross-device tracking, or fingerprinting of encrypted network traffic, which has been demonstrated for websites, apps, and voice commands. Understanding how these attacks work and how effective they are is a crucial precondition for the next steps: designing effective countermeasures and protections.

2) Detection of AI-based attacks. Detecting ongoing attacks and correctly attributing their source is particularly important for attacks that target humans, such as misinformation or phishing, which rely on AI-generated content. If we can attribute this content to AI in general, or even to the specific model that powers the attack, we can create awareness-based protections, such as labels and warnings in user interfaces, that may be easier to deploy than technical countermeasures that aim to prevent the attack. At the same time, awareness-based protections are desirable because they empower users to defend themselves. Existing approaches for detection rely on AI to label content as real or generated [226], however, it is not clear how futureproof and generalizable these approaches are.

3) Quantification. The effectiveness of attacks and defenses is commonly quantified with traditional machine learning metrics such as precision and recall. While these metrics are useful to compare the effectiveness of new attacks/defenses with existing attacks/defenses in controlled experimental settings, they are not sufficient to evaluate the effects in realistic settings. Additional metrics are needed, for example, to evaluate the feasibility of attacks, to estimate economic effects to inform regulators and lawmakers, or to quantify population-level privacy harms.

Statement from Expert 12: *1) Research incentives.* Research on offensive AI traditionally focuses on exploring the potential of AI to automate and enhance a wide range of malicious processes. These include vulnerability discovery, exploit development, penetration testing, evasive malware generation, encrypted traffic and hardware side-channel analyses, device fingerprinting, inference of private data, and more. While scientific articles exploring and prototyping these attacks are abundant, the impact of the research direction as a whole may be threatened by certain common practices. The field has largely stepped away from the underlying goal of offensive research – understanding and mitigating potential threats –, and instead centers around advancing simulated attacks until they surpass state-of-the-art performance. While the technical novelties behind AI-based attacks can be impressive and educating, aiming exclusively for superior attack performance may have harmful consequences on science. There is a tangible risk that research that does not delve into the exact workings of the novel attack vector to sufficiently inform mitigations does little to progress the field beyond the existing knowledge base. If offensive capabilities outpace defensive measures in research, this may inadvertently encourage a competitive mindset that prioritizes breaking systems over securing them, leaving systems chronically more vulnerable overall. To collectively address this open problem, it is imperative for the community to consolidate new practices, where an in-depth understanding of the novel AI-based attack mechanism and suggesting (and possibly evaluating) potential mitigations receive more attention and value (instead of being perceived as a limitation of the attack). Note that while this concern affects the entire offensive security field, it is especially challenging in the context of AI, where data-driven attacks are automated and inherently opaque, demanding targeted explainability measures.

2) Realistic attack simulation. In many offensive AI applications common evaluation practices have been established for the sake of simplicity, reproducibility, and compatibility with prior work. The consequences are as follows: (i) most of the studies do not or cannot aim for realistic estimations of threat severity, and (ii) the simplifying assumptions are often unacknowledged or inadvertently omitted, thus hindering the objective assessment of results. To yield results that more accurately reflect real-world or that can be more reliably assessed, the simplifying assumptions behind the data, the chosen metrics and the evaluation settings, as well as assumptions behind baseline attacker capabilities, need to be gradually overcome and systematically acknowledged.

1) Threat of offensive generative AI (GenAI). Offensive GenAI is a novel research direction that presents an unprecedented level of urgency due to the wide-spread adoption. The quality of AI-generated content has recently surpassed any expectations, demonstrating blasting performance against the weakest link in secured systems: humans. Social engineering, phishing attacks, spread of misinformation and fake content, and other kinds of AI-assisted manipulation demand focused attention from the scientific community. The core goals can be: (i) developing research practices that allow reliable simulations of the novel threat to reflect real-world infrastructures; (ii) encouraging interdisciplinary collaborations with non-technical fields (policymakers, educators, sociologists, psychologists, ethicists, etc.) for correct study designs and accurate interpretation of human behaviour; (iii) directly utilizing findings from offensive GenAI research to inform not just automated defenses, but, crucially, effective awareness campaigns.

In summary, conscious and effective design of defenses is a shared ultimate goal of offensive and defensive research teams and relies on raising our standards when harvesting solid insights from AI-based attacks.

B. Systematic Analysis via NLP: methods and results

We provide low-level technical detail (as well as the complete results) of the natural language processing (NLP) methods we used to analyze the expert statements.

1) ***N-grams analysis****:* As a preliminary check, we process the entire statements and extract the 20 most common bigrams/trigrams. The results are as follows: as we expected, the most common n-gram is “offensive ai” (61 occurrences); the second most common is “use ai” (11 occurrences), the third is “generative ai” (10 occurrences), followed by “social engineering,”, “ai generated” and “ai based” (9 occurrences each). Then, we have “real world” and “cognitive biases” (7 occurrences each). Next, with 6 occurrences, there are “threat actors”, “problems field”, “offensive ai tools”, “ai used”, “ai tools”, “ai systems”. With 5 occurrences, there are “open problems”, “generated content”, “benign ai”, “ai generated content”. Finally, with 4 occurrences, there are “social media” and “social engineering attacks”.

2) ***Keyword Extraction.****:* We then analyze each statement individually by extracting the most relevant keywords, using KeyBERT [208], a popular text-mining technique (used also, e.g., in [227]). Specifically, KeyBERT takes some text as input, and returns a list of keywords, each provided with a number (normalized between 0 and 1) representing its relevance in the text provided as input. Before running KeyBERT, we remove the common stopwords: [“offensive”, ‘‘ai”, ‘‘artificial”, “intelligence”, “malicious”, “attack”, “attacks”, “security”, “harmful”, “research”, “problems”]. The following are the top-5 keywords returned by KeyBERT for each expert statement (ES);

ES1: [(‘biases’, 0.2294), (‘dangers’, 0.2205), (‘discourse’, 0.2051), (‘increasingly’, 0.1979), (‘academic’, 0.1916)]
ES2: [(‘malware’, 0.304), (‘threat’, 0.296), (‘defense’, 0.2564), (‘detection’, 0.2564), (‘vulnerabilities’, 0.243)],
ES3: [(‘countermeasures’, 0.3488), (‘cyber’, 0.332), (‘tools’, 0.3243), (‘threats’, 0.3089), (‘tool’, 0.2919)],
ES4: [(‘technological’, 0.4223), (‘technology’, 0.4061), (‘exploitation’, 0.3831), (‘safeguards’, 0.3655), (‘exploiting’, 0.365)],
ES5: [(‘phishing’, 0.3615), (‘exploiting’, 0.3552), (‘biases’, 0.3445), (‘bias’, 0.3247), (‘behavioral’, 0.2992)],
ES6: [(‘attackers’, 0.3865), (‘malware’, 0.3828), (‘hackers’, 0.3294), (‘vulnerabilities’, 0.3008), (‘defending’, 0.2907)],
ES7: [(‘phishing’, 0.3917), (‘threats’, 0.3853), (‘deepfake’, 0.3744), (‘deepfakes’, 0.3721), (‘cyberattacks’, 0.3679)],
ES8: [(‘defense’, 0.4248), (‘defend’, 0.3772), (‘adversaries’, 0.3539), (‘defenders’, 0.3315), (‘mitigate’, 0.2803)],
ES9: [(‘attackers’, 0.4356), (‘threat’, 0.3449), (‘exploits’, 0.336), (‘malware’, 0.3313), (‘tactics’, 0.3311)],
ES10: [(‘disinformation’, 0.3484), (‘reality’, 0.3289), (‘surveillance’, 0.3198), (‘threats’, 0.3067), (‘attackers’, 0.2937)],
ES11: [(‘privacy’, 0.4319), (‘protections’, 0.3869), (‘phishing’, 0.373), (‘countermeasures’, 0.358), (‘defenses’, 0.3413)],
ES12: [(‘defenses’, 0.4028), (‘threats’, 0.3804), (‘exploit’, 0.3801), (‘malware’, 0.3642), (‘attacker’, 0.3562)]]

Our repository includes the code to generate this output [54].

3) ***Topic Modeling****:* Finally, we use topic modeling to extract the most relevant “topics” envisaged by our experts. To this end, we rely on BERTopic [209] (used in, e.g., [228]). Specifically, BERTopic takes as input a collection of documents, and returns as output a finite number of “topics:” each topic is a list of words which collectively represent a certain concept; each word of a topic has also a certain weigh in the overall definition of the overarching concept. To apply BERTopic, we proceed as follows:

- We take all the expert statements and split them into individual sentences, totaling in 218 sentences (5,082 words).
- After preprocessing these sentences (using [229]), we send them to BERTopic [209], specifying to output the 8 most relevant topics (we follow the default configuration of [209]).

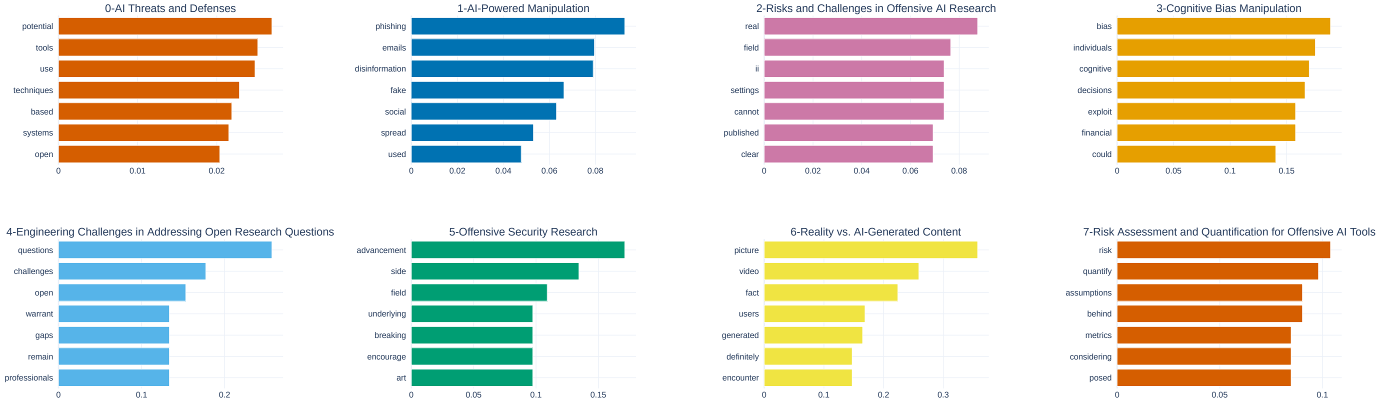
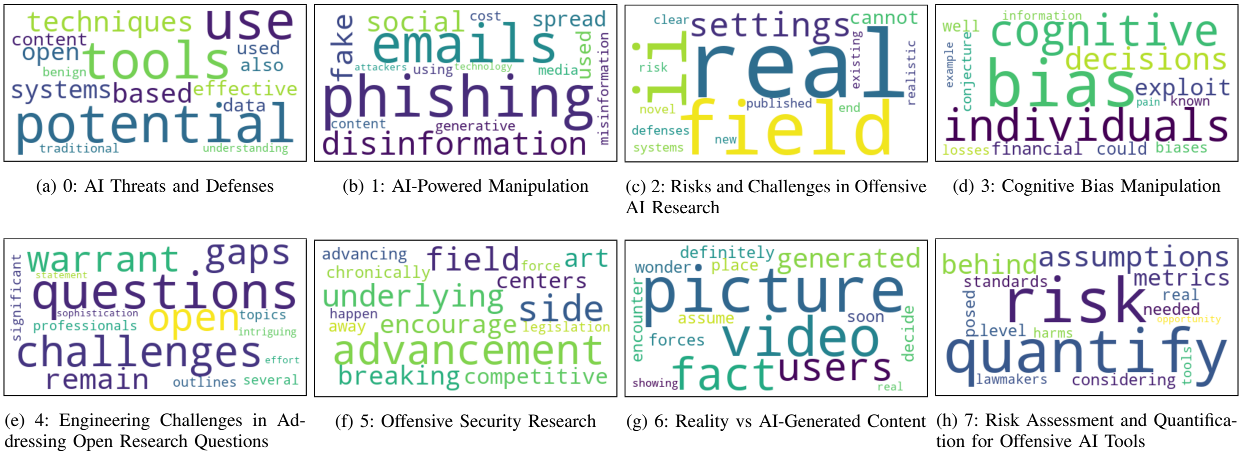


Fig. 12: **Topics identified by BERTopic.** We use BERTopic to analyze our expert statements and output the 8 most relevant topics. Each plot in the figure refers to a topic (title), wherein the y-axis shows the six most relevant words in the topic, and the x-axis denotes the weight of each word.



(a) 0: AI Threats and Defenses

(b) 1: AI-Powered Manipulation

(c) 2: Risks and Challenges in Offensive AI Research

(d) 3: Cognitive Bias Manipulation

(e) 4: Engineering Challenges in Addressing Open Research Questions

(f) 5: Offensive Security Research

(g) 6: Reality vs AI-Generated Content

(h) 7: Risk Assessment and Quantification for Offensive AI Tools

Fig. 13: **Word Clouds of the topics identified by BERTopic.** Each subfigure reports the word cloud of each of the 8 topics identified by BERTopic.

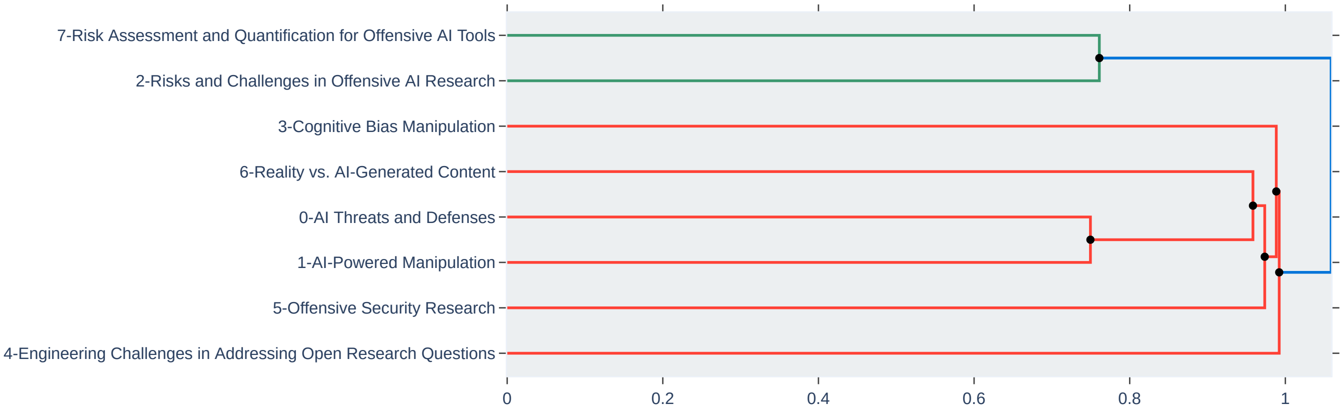


Fig. 14: **Hierarchical clustering of the topics identified by BERTopic.** We visualize the 8 topics identified by BERTopic to discern similarities.

- We take the 8 topics provided by BERTopic, remove noisy stopwords (the same we considered for the keyword extraction), and use LLAMA2 [230] to "label" each topic.

The results of these operations are summarized in Fig. 12, showing the 8 topics (labeled by LLAMA2) provided by BERTopic, alongside the weights of the six most-relevant words for each topic. We also report in Figs. 13 the word clouds defining each topic. Finally, we show in Fig. 14 the visualization of the 8 topics after having been analyzed via an hierarchical clustering algorithm (integrated in BERTopic [209]), allowing to discern how similar these topics are to each other. We provide the low-level source code of the abovementioned operations in our repository [54].

C. Authorship and Credits

This SoK is the result of a collective effort stemming from many individuals—including the 12 experts that provided their statements. However, the contributions of these 12 experts go beyond the mere 300–500 words paragraph outlining open problems on OAI—which, we remark, represents the basis of one of this SoK's major contributions (ℂ3).

First, the experts provided a first-round of feedback after reading our draft (which ended after the current Section V, and for which Section II was still incomplete) alongside providing their statements. No critical issues had been identified with our overarching goal and research methodology, but their remarks were invaluable in identifying and addressing some problems in the presentation and scope of our work. Then, the experts also provided a second-round of feedback on a revised version of our draft, which included the systematization of the experts' statements into the set of open problems and concerns on OAI (i.e., the current Section VI) as well as the discussion and conclusions (Sections VII and VIII). The experts' remarks also encompassed the appendices. All of these contributions improved the quality of our SoK tremendously. In addition, the experts also contributed by: *(i)* assisting in the revision of our SoK after receiving the reviews for another version of this work submitted to another top-tier security venue; *(ii)* approving the "submitted' version of this paper for SaTML25; and *(iii)* assisting in the rebuttal phase of SaTML25, including the preparation of this "revised" version of our SoK.

Due to the above mentioned reasons, the 12 experts have been listed among the co-authors of this paper. Every co-author of this work fulfills the authorship criteria embraced by IEEE [231]. For transparency, we provide below our CrediT statement [232], outlining how each co-author contributed to this SoK; we also report the full details of each author (some of which were omitted from this paper's first page).

- **Saskia Laura Schröer** (*saskia.schroeer@uni.li*): Conceptualization, Methodology, Software, Formal analysis, Validation, Investigation, Data Curation, Writing (Original Draft, Review, Editing), Visualization.
- **Giovanni Apruzzese** (*giovanni.apruzzese@uni.li*): Conceptualization, Methodology, Validation, Visualization, Writing (Original Draft, Review, Editing), Supervision
- **Soheil Human** (*soheil.human@wu.ac.at*): Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing (Original Draft, Review, Editing), Visualization
- **Pavel Laskov** (*pavel.laskov@uni.li*): Conceptualization, Formal Analysis, Resources, Writing (Original Draft, Review, Editing), Validation, Supervision, Project Administration, Funding acquisition
- **Hyrum S. Anderson** (*hyrum@robustintelligence.com*): Writing (Review & Editing), Validation
- **Edward W. N. Bernroider** (*edward.bernroider@wu.ac.at*): Writing (Review & Editing), Validation
- **Aurore Fass** (fass@cispa.de) [affiliated with CISPA Helmholtz Center for Information Security]: Writing (Review & Editing), Validation
- **Ben Nassi** (*nassiben@technion.ac.il*) [affiliated with Technion - Israel Institute of Technology]: Writing (Review & Editing), Validation
- **Vera Rimmer** (*vera.rimmer@kuleuven.be*) [affiliated with DistriNet @ KU Leuven]: Writing (Review & Editing), Validation
- **Fabio Roli** (*fabio.roli@unige.it*): Writing (Review & Editing), Validation
- **Samer Salam** (*ssalam@cisco.com*): Writing (Review & Editing), Validation
- **Ashley Shen** (*ashlshen@cisco.com*): Writing (Review & Editing), Validation
- **Ali Sunyaev** (*sunyaev@kit.edu*): Writing (Review & Editing), Validation
- **Tim Wadhwa-Brown** (*twadhwab@cisco.com*): Writing (Review & Editing), Validation
- **Isabel Wagner** (*isabel.wagner@unibas.ch*): Writing (Review & Editing), Validation
- **Gang Wang** (*gangw@illinois.edu*) [affiliated with the University of Illinois Urbana-Champaign]: Writing (Review & Editing), Validation

In the authors' list, the 12 experts have been alphabetically ordered.

Appendix E
Extra Information on our Research Methods

A. Papers and Briefings: comparison and discipline

We provide in Fig. 15 the yearly distribution of the academic papers and InfoSec briefings over time.

Moreover, we find it instructive to analyze the *discipline* of the venues of each work included in our literature systematization. Indeed, recall that our literature search encompassed repositories (§II-A) of a wide range of scientific disciplines. It is hence insightful to highlight such a diversity—especially given that OAI is a theme that can be tackled from diverse perspectives. To infer the discipline of each venue in an objective way, we relied on the Scimago database [233]: by querying this database with the name of a given venue, the database returns metadata of such venue—including its "subject area and category," such as Computer Science (CS) or Engineering. We hence query this database for each venue of the 95 papers included in our literature systematization, and report the "primary" subject area and category according to Scimago. The results are in the "Discipline" column in Tables I and II. We make some remarks.

- *Multi-disciplinary works.* Some venues in the Scimago database are associated with more than one subject area. For instance, "Applied Sciences" (the venue in which, e.g., the paper by Yu et al. [68] was published) is associated to Engineering, Physics, and CS (among others). In these cases, we infer the most appropriate discipline depending on the corresponding paper. E.g., for the work by Yu et al. [68], we assigned CS since it was the closest match.
- *Indexing.* The Scimago database is large and extensively curated, but some venues are not indexed. For instance, we could not find any entry for the "European Interdisciplinary Cybersecurity Conference" (which is the venue of [81]). In these cases, we assigned a discipline after qualitatively scrutinizing the works published in such venue. Nonetheless, lack of indexing should not be taken as poor value of any work in our systematization: first, because some venues are

still supported by reputable scientific organizations (e.g., the “European Interdisciplinary Cybersecurity Conference” is affiliated to the ACM [234]); second, because 94 out of 95 papers have passed peer-review (the only ones for which no published version exists are the paper by Tran et al. [106], which currently has over 45 citations on Google Scholar; and the one by Toemmel [105]).

- *Ranking.* For most venues, Scimago also provides a ranking, expressed in terms of “quartiles” (e.g., Q1–Q4, with Q1 being the highest rank). We preferred not to report this information in our Tables. First, because venue ranking is not necessarily proof of a paper’s quality (there are fundamental flaws even in papers accepted to IEEE S&P [62]). Second, because it would be misleading since some disciplines may have different rankings: for instance, the work by Iqbal et al. [79] was published in “Frontiers in Communication and Networks”, which is a journal whose primary field is CS, but for which the ranking is provided for two categories of CS (Q1 for “Computer Networks and Communications” and Q2 for “Signal Processing”). Third, because some venues (especially conferences) do not have any quartile: for instance, the work by Anand et al. [126] was published in CODASPY’18 which is not ranked (an alternative would be to use other rankings, e.g., CORE, but this would create issues for multidisciplinary venues). Therefore, we preferred not to provide any ranking-related information in our Tables.

By observing the Disciplines in Table I, we see that most works are from CS: only one is from Economics [98], and two from Engineering [86, 102]. This is expected given that the papers in Table I are highly technical. In contrast, for non-technical papers in Table II, there are eleven works from CS, but four works are published in Social Sciences venues, and one even for Medicine [146]. Therefore, we endorse future research to build on our systematization by considering papers from diverse fields.

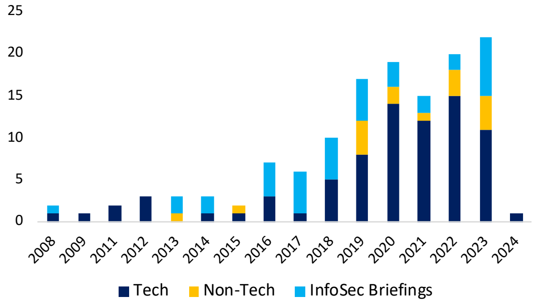


Fig. 15: **All works on OAI per year.** We present the yearly distribution of the works on OAI considered in this SoK, distinguishing technical and non-technical academic publications (§III) from InfoSec briefings (§IV).

B. History of the term “offensive AI”

In Figure 16 we provide a supplementary illustration of the different terms that have been used to refer to offensive AI.

C. Timeline of our research (and challenges)

To realize our SoK, we carried out diverse research activities. We provide a timeline in the Gantt chart shown in Fig. 17. In what follows, we will describe the temporal evolution of our activities, describing also some challenges we encountered.

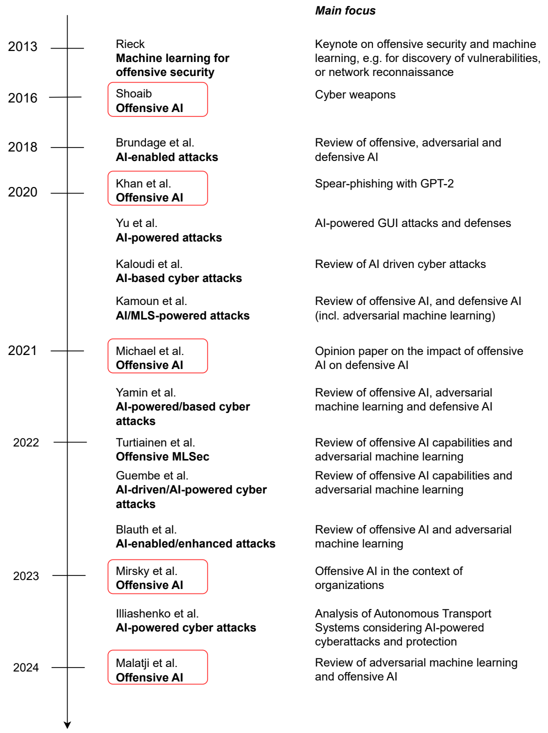

Main focus
2013
Rieck
Machine learning for offensive security
Keynote on offensive security and machine learning, e.g. for discovery of vulnerabilities, or network reconnaissance
2016
Shoaib
Offensive AI
Cyber weapons
2018
Brundage et al.
AI-enabled attacks
Review of offensive, adversarial and defensive AI
2020
Khan et al.
Offensive AI
Spear-phishing with GPT-2
Yu et al.
AI-powered attacks
AI-powered GUI attacks and defenses
Kaloudi et al.
AI-based cyber attacks
Review of AI driven cyber attacks
Kamoun et al.
AI/MLS-powered attacks
Review of offensive AI, and defensive AI (incl. adversarial machine learning)
2021
Michael et al.
Offensive AI
Opinion paper on the impact of offensive AI on defensive AI
Yamin et al.
AI-powered/based cyber attacks
Review of offensive AI, adversarial machine learning and defensive AI
2022
Turtiainen et al.
Offensive MLSec
Review of offensive AI capabilities and adversarial machine learning
Guembe et al.
AI-driven/AI-powered cyber attacks
Review of offensive AI capabilities and adversarial machine learning
Blauth et al.
AI-enabled/enhanced attacks
Review of offensive AI and adversarial machine learning
2023
Mirsky et al.
Offensive AI
Offensive AI in the context of organizations
Illiashenko et al.
AI-powered cyber attacks
Analysis of Autonomous Transport Systems considering AI-powered cyberattacks and protection
2024
Malatji et al.
Offensive AI
Review of adversarial machine learning and offensive AI

Fig. 16: **The evolution of the term “offensive AI.”** The term offensive AI (and similar related terms) has substantially evolved over time.

We began with a literature review—as is the case when approaching novel research directions. We started to do this in Summer 2023. During this investigation, we realized that the term “offensive AI” was not widespread in the literature, and that many papers proposed AI applications that could be used offensively—but did not use the term “offensive AI” (or a derivative) in the paper. Such preliminary findings prompted us to adopt a more systematic approach which entailed a qualitative assessment: after identifying a set of candidate papers, we would scrutinize such papers to determine if they fell in our own definition of the term “offensive AI.” As an additional sanity check, we have also tried to replicate the procedure followed by some prior work [248], i.e., by looking only at papers published in top-tier venues—intent in finding the coverage of OAI in this select number of venues. To this end, we considered all publications in the top-4 security conferences (NDSS, CCS, S&P, USENIX SEC) between 2021–2023, and we carried out a simple keyword search. Specifically, we looked for any full paper that had the term “offensive AI” / “offensive artificial intelligence” / “offensive machine learning” / “offensive ML” in either the abstract or title. We found only one paper that matched this criteria [249]. We believe this number underestimates the real number of publications discussing Offensive AI (as evidenced, e.g., by Mirsky et al. [20]). Therefore, such a finding confirms that our choice of *(i)* carrying out a larger search (encompassing four popular scientific repositories) that considers “any” type of peer-reviewed work (irrespective of the “ranking” of

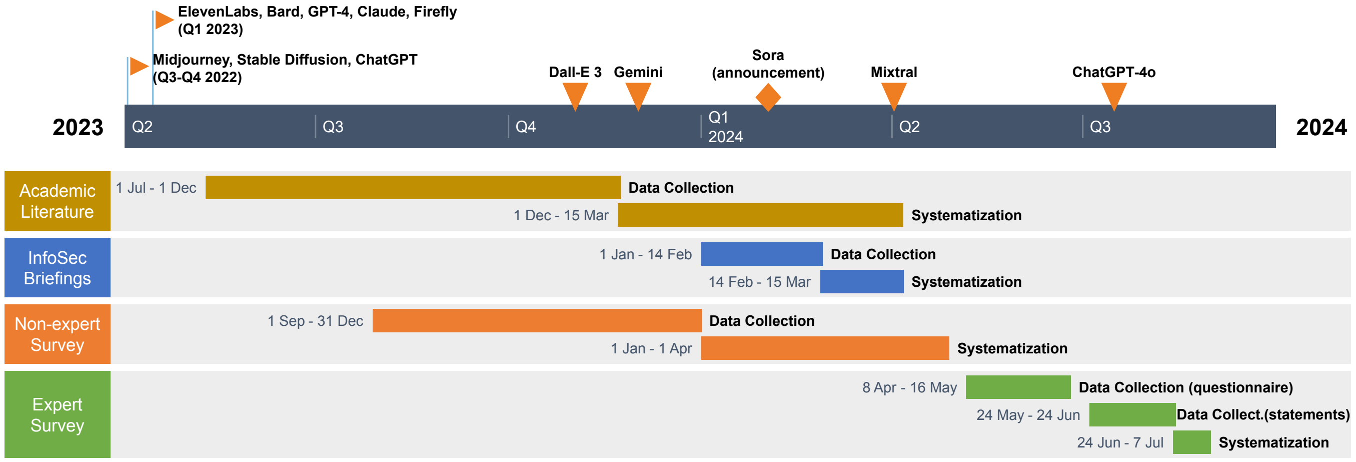


Fig. 17: **Timeline of our Research.** We began working on this SoK at the beginning of 2023. Throughout the entire activities shown in the figure, we have also had frequent meetings to steer the direction of our research and also to revise the paper. For instance, gaps in the "expert survey" swimlane are due to the necessity of analysing the experts' initial input before giving them the draft of our SoK. The timeline also shows important milestones in the context of OAI, represented by the (public) release of popular large language models that can process text, audio, images, or video content. Sources: [235–247]

any given venue); and then *(ii)* qualitatively analyzing each returned paper to determine if it fell in our definition of OAI, to be appropriate for our goals. Notwithstanding, the data collection phase of our literature analysis (for which we considered papers matching our queries and taken from IEEE Xplore, the ACM DL, arXiv, and Google Scholar) spanned between July and December 2023: during this timeframe we also progressively screened the papers returned by our search queries. We ultimately stopped our literature search, in December 2023, at which point we obtained a set of 95 papers that fell in our definition of OAI. We began systematizing these papers with the goal of identifying relevant aspects related to OAI—captured by our checklist; this analysis required several months of work by multiple authors, and terminated in March 2024.

The second major activity we have carried out was the user study with non-experts. The idea of carrying out this study was inspired by the initial findings of our literature analysis. Specifically, we conjectured that the potential applications of OAI were so diverse that it could be insightful to ask "non-experts" about their opinion on OAI. We designed our questionnaire and began disseminating it in September 2023. We stopped collecting responses at the end of December 2023. We then analyzed the collected responses and derived our codebook (which we knew would be used also later for analyzing the experts' input).

The third activity we have carried out is the analysis of InfoSec briefings—which spanned between Jan. and March 2024. We posited that these venues could provide a complementary perspective on the "practical" use cases of OAI in the real world—especially given that not many research papers showcased real-world demonstrations of OAI. This procedure was not trivial—despite the existence of a much lower number of InfoSec briefings than research papers. Indeed, while *finding* the briefings was simple, *ascertaining* whether a briefing is about OAI required us to watch the entire video (∼30–40m long) of the presentation, in some cases. Moreover, to systematically analyze the OAI-related briefings, we also had to rely on the video. This may explain why considering InfoSec briefings is uncommon in the SoK literature.

The last activity we have carried out is the user study with experts. First, we stress that most of the experts we contacted accepted our request to participate in our research. After finding an agreement with 12 experts, we distributed our questionnaire; we did this in April 2024. Then, in the following weeks, we analyzed the responses we collected and also assembled what was the initial draft of our SoK. We shared this draft with the experts at the end of May 2024, and gave each expert 3 weeks of time to provide their statement of 300–500 words (one expert was late and submitted their contribution 1 week later). At the end of June 2024, we received all the statements and systematically analyzed them. We then shared an improved version of our SoK with the experts so that they could provide feedback—which was invaluable to clarify misunderstandings and improve the clarity of our work.

Real-world developments. In Fig. 17, we have also reported some major real-world events in the context of OAI. Specifically, we consider that *the public release of LLM*, which undoubtedly enable anyone (including evildoers) to use AI, represent important milestones. We observe that *all our research activities have been carried out well after* the public release of powerful LLMs (such as Bard, ChatGPT, Claude, Stable Diffusion). Therefore, we find it unlikely that, e.g., the rollout of Dall-E3 or Gemini may have impacted the responses of our non-expert survey; at the same time, the release of ChatGPT-4o is also unlikely to have substantially affected the ideas of our experts. Ultimately, the AI field is very fast paced, and educated people are likely aware of this fact.