"Hey Google, Remind me to be Phished" Exploiting the Notifications of the Google (AI) Assistant on Android for Social Engineering Attacks

Marie Weinz, Saskia Laura Schröer, Giovanni Apruzzese
Liechtenstein Business School, University of Liechtenstein
{marie.weinz, saskia.schroeer, giovanni.apruzzese}@uni.li

***Abstract*—We showcase how to maliciously exploit a functionality of the Google ecosystem (specifically, of Android) by elucidating how the notifications generated by the Google Assistant may help phishers in reaching their goals. We found that Android users who have Google Assistant check their inbox will be reminded to carry out duties that are solicited in emails that *have never been opened before*. From a social-engineering perspective, attackers can send specific emails to Android users, and these users will receive notifications (from Google) "reminding" them that a task is soon due, thereby urging them to "fall for phish." Just imagine: while going through your day, you suddenly receive a notification on your smartphone saying that "An outstanding task is soon due." Tapping on the notification leads to opening an email which, if malicious, contains ill-purposed content, such as harmful links or malware attachments. The sense of urgency from the unexpected reminder may lead to overlooking some phishing cues—facilitating social engineering attacks.**

This subtle (and novel) threat is rooted in the quintessential functionalities of smart (AI-based) assistants that passively analyze our data to improve our digital well-being. Users of these tools must be made aware of this issue to prevent harmful consequences. Therefore, besides describing our discovery and analysing it under a security lens, we also (i) carry out a user study to gauge the potential impact of this issue; and (ii) emphasize some practical takeaways for both users and developers. We disclosed our finding to Google: they acknowledged the possibility of attacks, but stated that no fix to their software will be made.

I. Introduction

With ∼4 billion users worldwide [1], Android is the leading operating system (OS) of modern smartphones (71% market share [2]). Thanks to its integration with the Google's ecosystem (e.g., GMail), owners of Android devices can benefit from the continuous updates made by one of the world's top tech companies [3]. Among the most recent developments that have substantially enhanced the quality of experience of Android users, the *Google Assistant* stands out [4]. Powered by artificial intelligence (AI) [5, 6], the Google Assistant monitors the plethora of activities that its users carry out during their daily digital lives—providing tools and resources (e.g., automatic reminders [7]) that improve the users' overall well-being [8].

Unfortunately, such a large reservoir of users makes the Android ecosystem an attractive target for cyberattackers—and, in particular, for *phishers* [9–12]. Indeed, some specific functionalities of Android OS, such as its notification system, can be maliciously exploited to facilitate social engineering attacks—and some of these "security vulnerabilities" have been discussed in prior works [12]. In this paper, we present a novel way through which *the Google Assistant can be leveraged for social engineering attacks* against Android users.

How does it work? Our attack is rooted on the hypothesis that Google Assistant *perpetually checks the inboxes* of the email accounts associated to Android smartphones. This serves to "help" users, so that if they receive an email stating that, e.g., "a task is due soon", *a notification will be triggered* on their smartphones to warn them. However, the problem is that the Google Assistant "blindly trusts" the analysed emails—including those concealing social engineering attempts (and, unfortunately, existing phishing email filters can be trivially bypassed [13]). An attacker can exploit such a functionality to carry out phishing campaigns, i.e., by using the automatic reminders of Google Assistant as a catalyst to instill a sense of urgency [14] in their victims—who may be more likely to open the email and, e.g., click on a malicious link (Fig. 1).

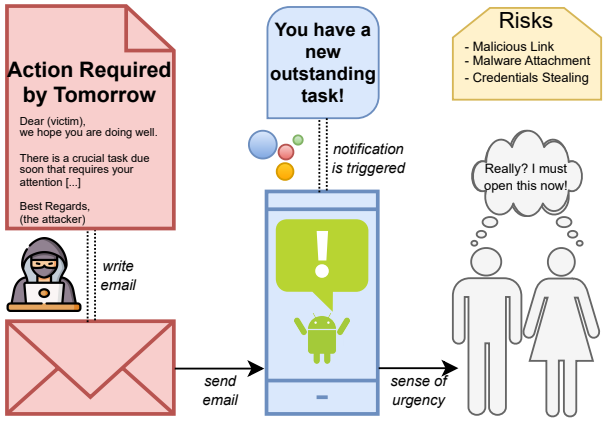


Fig. 1: **Leveraging the Google Assistant for Social Engineering.** An attacker writes an email stating: "Action required soon". The email will trigger a notification from the Google Assistant (within Android), which will remind its users of an outstanding task. The users (i.e., the victims), driven by the sense of urgency, may carelessly open the email and fall for a phishing trap.

We will verify our hypothesis with practical experiments, and examine what may be done by Android smartphones behind the back of its users—who we found may not be very knowledgeable of the Google Assistant. Yet, we also provide another "meta" contribution: we will *tell a story* explaining how we discovered the issue described in this work, which happened by pure chance. This serves to highlight that anyone could have come to the same conclusion—including attackers.

CONTRIBUTIONS. We raise awareness on a subtle issue that can propel phishing activities across Android. Specifically, we:

- discover a **way to maliciously exploit a helpful functionality of Android** which facilitates social engineering by *leveraging the notifications* sent by the Google Assistant (§II);
- validate our finding (§III) and disclose it to **Google who acknowledged that our discovery** can result in social engineering attacks, but *refused to apply any fix* (§IV);
- shed more light on our discovered issue (§V), gauge its potential impact through a user study (§VI), and *provide recommendations* to mitigate its effectiveness (§VII).

At the point of writing this paper, the problem has not been fixed yet. We recorded a video showcasing an end-to-end workflow of our attack (provided in our repository [15]).

Privacy Notice. To provide evidence of our discovery, we will show images capturing confidential details of the authors. To protect our privacy, some elements are obscured (in black).

II. DISCOVERY ("IT WAS JUST ANY OTHER DAY WHEN…")

We present our "attack" by narrating a story. Specifically, we describe the interactions between the two individuals who brought this issue to light. These individuals are a Marie, 𝕄, and Giovanni, 𝔾. In what follows, we explain how 𝔾 and 𝕄, while working side-by-side on a research project focused on phishing education, realised that the Google Assistant can be leveraged for malicious purposes. To better convey the role of "daily routines" in the process of discovering security problems, the following content is written in a relaxed tone.

A. Backstory (why did we even stumble upon this?)

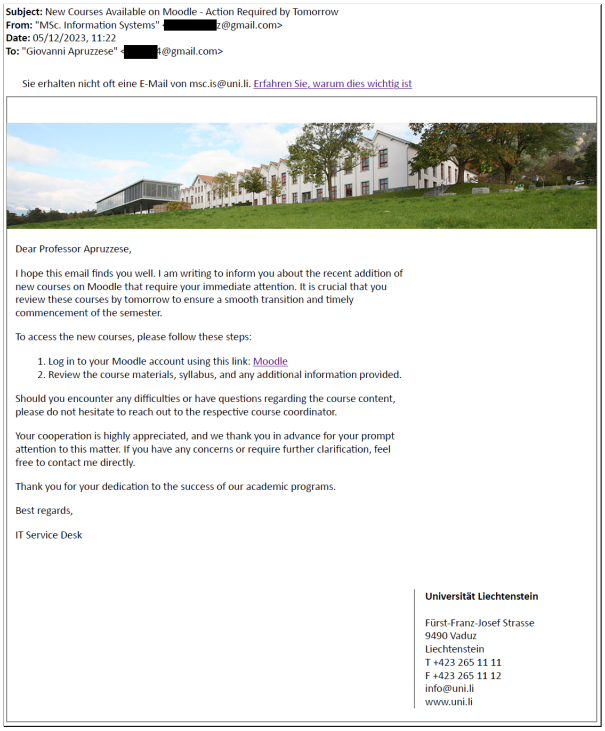
In September 2023, 𝔾 and 𝕄 had a meeting wherein they discussed the goals of the underlying research project: investigating the phishing education in modern organizations. Specifically, 𝕄 was going to carry out some "phishing-email training exercises" in several companies—under the supervision of 𝔾.

In October 2023, 𝔾 and 𝕄 had another meeting: 𝕄 found agreements with some companies, and 𝔾 suggested that state-of-the-art solutions to accomplish their goals were *GoPhish* [16] and *Zphisher* [17] (used, e.g., in [18, 19]). Indeed, GoPhish allows to craft phishing emails in bulk and, if combined with Zphisher, it also allows to embed customised links to determine whether such emails are read by its recipients, and log corresponding details (e.g., the IP addresses of the devices that clicked on the link).

In late-November 2023, 𝔾 and 𝕄 had another meeting. Given the sensitivity of the subject, and also given that neither 𝔾 nor 𝕄 had used GoPhish or Zphisher before, 𝔾 and 𝕄 agreed that 𝕄 was going to deploy an instance of these tools on their premises, and then carry out some pilot tests by sending some "phishing" emails to 𝔾. Specifically, the goal of these preliminary assessments was to ensure that such (simulated) "phishing" emails would not be blocked by the automatic spam filters that protect the (many!) email accounts of 𝔾. Then, if such emails were not blocked, the following step was to study what information from 𝔾 was "captured" by the considered tools (managed by 𝕄). All these operations were necessary to ensure that the overarching research (unrelated to the issue discussed in this paper—of which we were still not aware!) was carried out fairly, ethically, and with scientific rigour.

B. Realization ("wait, this is weird...")

On December 3rd 2023, 𝔾 received a "suspicious" email on one of their accounts (**4.gmail.com*). However, 𝔾 did not know (yet) the contents of the emails crafted by 𝕄; furthermore, 𝔾 provided five email accounts that 𝕄 was supposed to test. Hence, on December 4th, 𝔾 inquired 𝕄 whether the email was truly from 𝕄 (which was true), and also to send emails to each of the four other email accounts owned by 𝔾. The following morning, i.e., on December 5th, 2023, at 11:12AM (all times are CET), 𝕄 wrote a message to 𝔾 stating "Alright, I will send you more emails". Soon after, 𝕄 sent various emails—all having the same text, but sent to the five addresses specified by 𝔾. We report one of these emails in Fig. 2: this email is sent to **4@gmail.com* (owned by 𝔾), and is sent from **z@gmail.com* (owned by 𝕄) but whose name is spoofed to "MSc. Information Systems" (which is related to 𝔾's professional activity). The subject is "New Courses Available on Moodle - Action Required by Tomorrow", and the text describes various tasks that the recipient (i.e., 𝔾) was supposed to carry out for their job, and it contains a link ("Moodle", which leads to the webpage in Fig. 16) bound to Zphisher. The email is sent at 11:22AM. However, at that point in time, 𝔾 was still sleeping (this fact plays a crucial role).


Subject: New Courses Available on Moodle - Action Required by Tomorrow
From: "MSc. Information Systems" <[redacted]z@gmail.com>
Date: 05/12/2023, 11:22
To: "Giovanni Apruzzese" <[redacted]4@gmail.com>

Sie erhalten nicht oft eine E-Mail von msc.is@uni.li. Erfahren Sie, warum dies wichtig ist

Dear Professor Apruzzese,

I hope this email finds you well. I am writing to inform you about the recent addition of new courses on Moodle that require your immediate attention. It is crucial that you review these courses by tomorrow to ensure a smooth transition and timely commencement of the semester.

To access the new courses, please follow these steps:

1. Log in to your Moodle account using this link: Moodle
2. Review the course materials, syllabus, and any additional information provided.

Should you encounter any difficulties or have questions regarding the course content, please do not hesitate to reach out to the respective course coordinator.

Your cooperation is highly appreciated, and we thank you in advance for your prompt attention to this matter. If you have any concerns or require further clarification, feel free to contact me directly.

Thank you for your dedication to the success of our academic programs.

Best regards,

IT Service Desk

Universität Liechtenstein

Fürst-Franz-Josef Strasse
9490 Vaduz
Liechtenstein
T +423 265 11 11
F +423 265 11 12
info@uni.li
www.uni.li

Fig. 2: **The triggering "phishing" email.** The email is sent by 𝕄 to the personal email account of 𝔾. The email has "action required" in the subject, and mentions various tasks (related to 𝔾's job) to be completed by the next day. This email triggered a notification on 𝔾's smartphone from the Google Assistant. (Fig. 15a shows the inbox of 𝔾 when receiving the email.)

At ≈11:50AM, 𝔾 woke up. After picking up their smartphone (a Samsung Galaxy S23, running Android 13), 𝔾 began

going through all the notifications that had been generated during the morning. Obviously, such notifications were related to the personal and professional life of 𝔾. We report a screenshot showing such notifications in Fig. 3. At the top of this screenshot, there is the notification (in green) showing the message sent by 𝕄 (at 11:12AM). Then, there are other notifications—which are typical for 𝔾's routine. However, there was one notification that caught 𝔾's attention, circled in red in Fig. 3. Apparently, "Google Assistant", at 11:22AM (i.e., 10 minutes after the message from 𝕄, and at the same time as the email sent by 𝕄 in Fig. 2), stated that there is a task "Due by December 6th" (i.e., the following day). This prompted a reflection by 𝔾: according to 𝔾's memory, there was no task[1] due for either December 5th or December 6th. So *why was such a notification displayed on 𝔾's smartphone?*

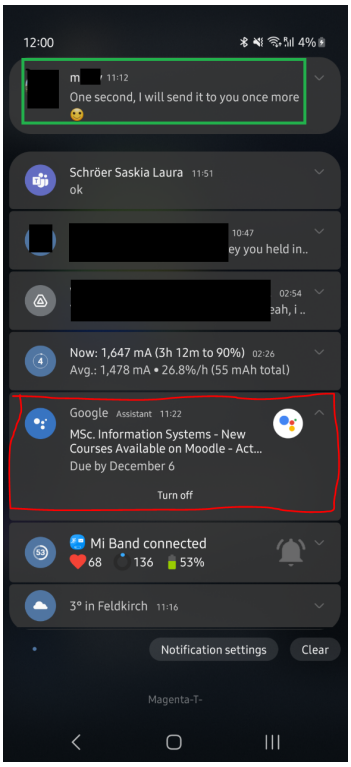


Fig. 3: **The weird notification of the Google Assistant.** We show (some of) the notifications displayed on the smartphone (Samsung S23, running Android 13) of 𝔾 in the morning of December 5th, 2023. The green rectangle highlights the message sent by 𝕄. The red rectangle shows the notification generated by the Google AI Assistant in response to the email sent by 𝕄 (shown in Fig. 2).

Recall that 𝔾 had just woken up and, hence, was still overwhelmed by the huge number of notifications on their smartphone. Nevertheless, 𝔾 soon realised that the notification could be related to the email sent by 𝕄: After all, the content of such notification (i.e., "MSc. Information Systems - New Courses Available on Moodle - Act...") resembled the content of the email that 𝕄 sent to 𝔾 on December 3rd.[2] Therefore, 𝔾 quickly dismissed the notification as being junk. Yet, 𝔾 felt that such a notification was "wrong:" granted, it originated from a fake email—but **why did such an email induce the Android OS of 𝔾's smartphone to generate a notification in the first place?** Indeed, 𝔾 tried to remember whether 𝔾 gave any consent about Android OS inspecting their inbox and generating notifications on the basis of the received emails—but 𝔾 could not remember ever giving such a consent (nor whether such an option even exists in the first place). Hence, we realised that this property can give rise to the following

⚠ **Malicious exploit:** "What if such *unsolicited notification* is maliciously exploited by attackers?" An evildoer can *(i)* send a phishing email to a user of an Android smartphone. Upon receiving the email, Google Assistant would *(ii)* instantly generate a notification which would "urge" the user to take action.[a] Driven by such a sense of urgency, the target would *(iii)* open the email[b] and, potentially, click on a malicious link, or carelessly share credentials (see Fig. 1).

[a] After all, the task explicitly said "due by tomorrow"!
[b] Tapping the notification brings to the email (see Fig. 13 in Appendix A)

III. Validation (is it just a "false positive" or...)

We have validated our discovery with further tests. Specifically, we verify whether it affects other devices (§III-A) and also if it has been patched by Google (§III-B). We aim to answer the question: "is our finding really a problem—today?"

A. First check: different devices and OS version

The events discussed in §II only show that our discovery affected a Samsung Galaxy S23 running Android 13 on December 5th, 2023. Hence, to verify whether this problem was not just a spurious and transient artifact, we repeated our experiment two weeks later.

Setup. On December 20st, 2023, 𝔾 asked 𝕄 to send the exact same emails once more. Accordingly, 𝕄 sent these emails to the five email addresses of 𝔾 (including **4.gmail.com*) at around 6:11PM. A few hours later, 𝔾 checked the notifications generated by two (physical) smartphones owned by 𝔾: the Samsung Galaxy S23 with Android 13 (used in the main discovery §II-B); and a Samsung Galaxy S10e with Android 9.

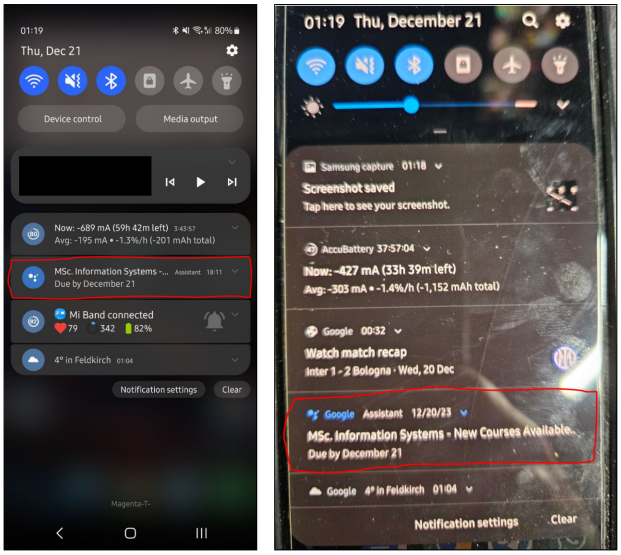
Results. We report the notifications shown by these two devices in Fig. 4. These figures, taken at 1:19AM of December 21st, 2023, show that both of these smartphones (*despite running different versions of Android*) are affected by the issue discussed in this paper. Indeed, the email "urging the recipient to take action before the following day" prompted both of these devices to generate a notification that could give birth to a social engineering attack. Moreover, the fact that the notification appeared another time is evidence that the problem is systematic of the Google Assistant—and that our finding was not just a random occurrence.

B. Second check: some months later (and different email)

After our first validation, we disclosed our findings to the Google team (on December 29th, 2023): we will discuss this in the next section (§IV). However, here, we discuss the results of our experiments after having repeated them three more times.

[1] Of course, 𝔾 checked their calendars: there were indeed no tasks "due by" Dec 5/6th (just "meetings" which had no specific requirement).

[2] **The notification also appeared for this email**. According to 𝔾, they dismissed the Dec. 3rd notification without noticing it. However, we later checked the notification history (see Fig. 15b) and confirmed this fact.



(a) **Samsung S23** (Android 13) (b) **Samsung S10e** (Android 9).

Fig. 4: **First Validation (Dec. 21st, 2023).** We repeat the experiment two weeks later: $\mathbb{M}$ sends the same email to $\mathbb{G}$. We report the notifications generated on the two Android devices owned by $\mathbb{G}$ (Fig. 4b was not tested in §II-B). Red boxes denote the notification generated by the email sent by $\mathbb{M}$.

Setup. To further validate and investigate our finding, we carry out the operations done for the previous experiment (§III-A) on March 10th, May 1st, and May 26th, 2024.

- on March 10th, $\mathbb{M}$ sends the same emails (sent in December 2023 from *z@gmail.com*) to the same email addresses of $\mathbb{G}$;
- on May 1st, we send the same email to the usual email address (**4.gmail.com*) but from *another* email address (also ending in *@gmail.com*). We also try different variants (e.g., by changing the email text and subject) to try to learn more about this issue (more details on this test are in §V-B).
- on May 26th, we make one last experiment in which we send a similar email, but from an email address that does not resolve to Google and which we created from scratch and for the specific purpose of this experiment (paul_reeves@onmail.com). The complete workflow of this experiment is shown in our demonstrative video [15].

In all cases, we then check whether such emails triggered the same notification in the Samsung S23 and S10e owned by $\mathbb{G}$.

Results. We report the screenshots (taken at 3:54PM of March 10th, 2024) of the first additional validatory experiment in Fig. 14 (in Appendix A). Both devices generated a notification for the email sent by $\mathbb{M}$ to the **4@gmail.com* email address of $\mathbb{G}$ (at 3:50PM), shown in a green rectangle. We also see, in a red rectangle, that both devices had the Google Assistant generate a corresponding "task is due soon" notification (at 3:51PM). With regards to the second validatory experiment (on May 1st), we also confirm that they triggered the exact same notification (the only difference is that the text of the notification does not say "MSc. Information Systems", but the name field of the different email address); we also found out that the notification is triggered by the string "Action Required" in the subject. Finally, for the last experiment (on May 26th), we confirmed that the notification is triggered even from a completely different email address. Hence, the problem had not been patched yet (despite having informed Google about this on December 28th, 2023).

⚠ **True Problem.** Our discovery is not a "false positive", has not been patched yet, and it affects (at least) devices running Android 13 and 9 (which cumulatively account for ≈30% of the overall distribution of Android OS [20]).

IV. Ethical Disclosure (what did Google say?)

For the sake of responsible disclosure, we contacted the Google Team and informed them of our discovery as soon as we became certain of it. Let us explain how this went by tracing a *timeline* of the evolution of our "submission".

- On December 29th (at 12:45PM), we made a submission (ID:318056254) to the Google's issue tracker [21].
- Four hours later, we received an update stating that our submission was taken into account and to respect confidentiality procedures. We promptly replied that we signed the Google "Contributor License Agreement" [22] and that we validated our discovery on two smartphones (describing the experiment in §III-A). We also exchanged two more messages, promising that we would not publicly announce this issue until April 2024.
- We received another update on January 2nd, 2024, stating that "Good news! According to Google magic, your report is likely actionable for us, so it has been moved up in our queue by raising the priority." This led to the bug being acknowledged as P2 priority and S2 severity; afterwards, the priority was changed to P3.
- The last update we received was on January 11th, 2024. The Google Team stated that "we've investigated your submission and made the decision not to track it as a security bug." The alleged motivation is that "the issue you're describing can only result in social engineering, and we think that addressing it would not make our users less prone to such attacks."

They also posted a link [23] explaining that "phishing or social engineering attacks" are typically not considered as security bugs by Google. Then they closed the issue, with motivation:

> **"Status: Won't Fix (Intended Behaviour)."** Factually, *Google acknowledged our discovery can be used to convey "phishing or social engineering attacks."*[a] However, Google decided not to fix it.[b] We agree that "fixing it" may require a huge amount of resources (and, potentially, a rethinking of the entire Google Assistant ecosystem). Yet, we argue that *something can be done* to "address" this problem—starting from responsibly informing users of its existence (§VII).
>
> [a] Such acknowledgement also implies that this issue also extends to other Android OS and smartphones (beyond those considered in this paper).
> [b] Such "Intended Behaviour" makes this issue even more subtle.

Nonetheless, we stress that the Google team was cheerful, and we respect their decision. Indeed, we do not want to point the finger at Google: it is true that social engineering can hardly be countered, and preventing all potential "phishing-exploitations" of Android may make Android unusable.

```
IP: 74.125.151.199
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/42.0.2311.135 Safari/537.36 Edge/12.246 Mozilla/5.0

IP: 74.125.151.200
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/42.0.2311.135 Safari/537.36 Edge/12.246 Mozilla/5.0

IP: 66.249.81.238
User-Agent: Mozilla/5.0 (Windows NT 5.1; rv:11.0) Gecko Firefox/11.0 (via ggpht.com GoogleImageProxy)

IP: 66.249.81.237
User-Agent: Mozilla/5.0 (Windows NT 5.1; rv:11.0) Gecko Firefox/11.0 (via ggpht.com GoogleImageProxy)

IP: 66.249.81.238
User-Agent: Mozilla/5.0 (Windows NT 5.1; rv:11.0) Gecko Firefox/11.0 (via ggpht.com GoogleImageProxy)

IP: 185 222
User-Agent: Mozilla/5.0 /537.36
```

Fig. 5: **The IP addresses that "checked" the email.** We inspect the logs of Zphisher to determine which hosts opened the email received by 𝔾.

V. Analysis (what is happening? workarounds?)

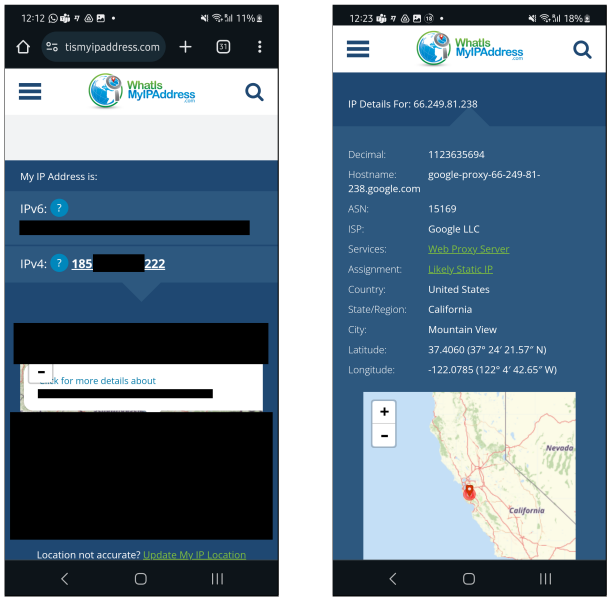
We further substantiate our contributions by carrying out additional analyses. First, we *attempt to explain* some peculiarities of this problem (§V-A). Then, we pose some *open questions* that we could not find an answer to (§V-B). Finally, we highlight the *poor transparency* of the Google Assistant to end users with respect to disabling its functionalities (§V-C).

A. Investigation: IP (and Email) Addresses

To shed further light, we analyse the logs of Zphisher and attempt to infer which email prompted the notification.

IP addresses. An instructive way to investigate what is done by the Android OS (and which leads to generating the notification) is to leverage the functionality of Zphisher [16] of logging the IP addresses that opened the link included in a given email. Hence, on December 5th, 2023, at 1:20PM, we checked the IP addresses of the devices that "read" the email delivered to 𝔾 and clicked on the link ("Moodle", leading to Fig. 16). We report such logging in Fig. 5, showing the IP addresses and user agents of each "visit" to this link; the results at the top are those which occurred earlier. Five IP addresses were logged by Zphisher: `74.125.151.199`, `74.125.151.200`, `66.249.81.238`, `66.249.81.237`, `185.?.?.222`. We first checked the public IP address of the devices in 𝔾's location (see Fig. 6a), confirming that the IP address pertaining to 𝔾 was the last one of the list captured by Zphisher: this makes sense, because 𝔾 did open the email eventually (proof is Fig. 13) and clicked on the link. Then, we checked the location of the four other IP addresses. First, we see that, despite being four, they are similar to each other. Hence, we lookup two of these IP addresses: `66.249.81.238` (shown in Fig. 6b) and `74.125.151.200` (shown in Fig. 7). Both of these IP addresses resolve to hosts of Google LLC, located in Mountain View, CA, USA—and they clearly have nothing to do with 𝔾 (who resides in a completely different geographical area).[3] Such a finding confirms that *devices owned by Google opened the email sent to 𝔾's email account and inspected the link* (and did not deem it as "suspicious").

Email addresses. As an additional analysis, we attempt to identify *which email* triggered the notification—and, hence, which email address was "checked" by the Google AI Assistant to generate the contents of the notification. Recall that, in some of our experiments (those in §II-A and the first validatory one in §III-A), 𝕄 always sent the emails to the five accounts owned by 𝔾, i.e., **4@gmail.com*, and four other ones—all being included in the GMail App installed on 𝔾's devices. Hence, we are still unsure if the notification is generated by "reading" the emails sent to **4@gmail.com* (which is the "primary" account of 𝔾 in their Android smartphones) or to any of the other four accounts. To this end, we perform another experiment: on March 17th, 2024 at 9:10PM, 𝕄 sends one email to **4@gmail.com*, without sending anything to the other four accounts. Such an email triggered the generation of the notification. Then, 𝕄 sent the email to the institutional email account of 𝔾: this mail was not received by 𝔾, but this mailbox is not managed by Google. Finally, 𝕄 sent the email to another GMail account of 𝔾: this email was received, but no notification from Google Assistant was generated. These results confirm that *the email address "read" by the Google AI Assistant is the one corresponding to the email account of 𝔾 associated to their Android devices*.



(a) **The (public) IP address of 𝔾** (b) **Lookup of `66.249.81.238`.**

Fig. 6: **Checking the IP addresses (Dec 5th, 2023).** We check the IP addresses logged by GoPhish (in Fig. 5). We first confirm that the last IP address corresponds to the (public) IP address of 𝔾. Then, we lookup the other IP addresses (Fig. 6b and Fig 7), and find they belong to Google.

[3] We also repeated such "tracking" on December 20th, 2023, and we found `66.249.81.129` and `66.249.81.128` also checked the email sent by 𝕄. Both of these IP addresses also belong to Google (located in Mountain View, CA).

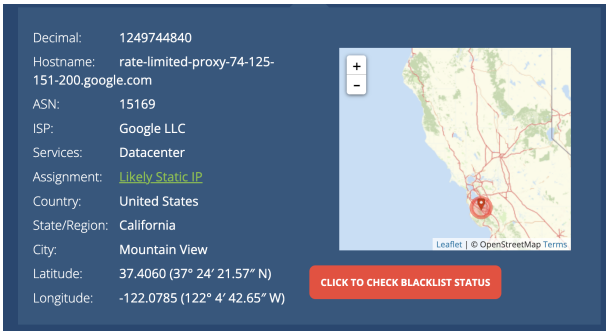


Fig. 7: **IP address lookup of `74.125.151.200`.** As in Fig. 6b, this IP address also belongs to Google (we do this lookup also on Dec. 5th, 2023).

B. Open Questions

We *cannot* be certain of what is done by Google (or Android) at the low-level. Furthermore, we *cannot* (nor seek to) reverse-engineer the Android OS so as to explain the ins-and-outs of our discovered problem: such an analysis would be unfeasible, since it may revolve around the intricacies of AI systems (which are currently poorly explainable [24, 25]). Here, we will list several "open questions" (supported by additional analyses) that underscore some privacy-related issues—which we will use to formulate our recommendations (§VII).

Why did hosts from Google appear in Zphish? To answer this question, we can conjecture two possible explanations—which are not mutually exclusive. **(1)** The most plausible explanation is that the contents of such emails were checked by Google's "spam filters." After all, the email account owned by 𝔾, **4@gmail.com*, is a GMail one. Yet, if the reason of such checks stems from antispam services, such services are not very effective: the email address used to send most of these emails, **z@gmail.com* (despite being also a GMail one), was very new; moreover, the content of such emails was unlikely to be generated by the owner of such an address: the name in the "from" field (i.e., "MSc Information Systems") was spoofed and had no connection with the actual email address (i.e., **z@gmail.com*). **(2)** Alternatively, it is possible that the visits are due to how the Google AI Assistant ecosystem works. For instance, this Android service may contact some server (owned by Google) to carry out a more detailed analyses of the payload of some emails. Such analyses are meant to provide more "accurate" results (w.r.t. those computed locally, on the Android device itself), usable to generate appropriate notifications[4] by the Google AI Assistant.

Why did the notification pertain to **4@gmail.com*? The notification, despite being clearly related to the *professional* life of 𝔾, was generated by an email sent to an account that revolved around the *private* life of 𝔾. Indeed, according to 𝔾, the emails received by **4@gmail.com* would hardly allow anyone to understand much about 𝔾's job.[5] It would have made more sense if, e.g., the notification stemmed from analysing the inbox of 𝔾's *institutional* email account. However, as we showed (in §V-A), this is not the case. Put differently: the notification generated by the Google AI Assistant pertains to an email mentioning details related to 𝔾's job, but the Google AI Assistant (in 𝔾's devices—which we proved analyses **4@gmail.com*) *should be oblivious* of such details. Hence, the notification should not have been generated in the first place—deeming such details as irrelevant (likely, spam). Indeed, according to 𝔾, nobody would send such "professional" emails to **4@gmail.com*. We find it surprising that the Google Assistant was not able to make such a logical connection (despite having access to the entire inbox of **4@gmail.com*).

What elements of the email triggered the notification? It is a fact that the notification[6] was created on the basis of the email received by 𝔾. Hence, we wonder what exactly led to that specific notification—and, in particular, what information the Google AI Assistant uses to generate the notification under scrutiny. We conjecture that such a process entails the following operations. First, the Google AI Assistant must *passively and continuously monitor the inbox of the user.* This is confirmed by the notification being generated immediately after 𝔾 received the email (see Fig. 14a). Then, the Google AI Assistant must *read at least the "from" and the "subject" of the email*. This is confirmed by the text of the notification being (almost) identical to these two fields (see Fig. 2 and Fig. 3). Finally, the Google AI Assistant performs some analyses (e.g., compare the date with the current day) to determine that the task is "due by" a certain day. Note that, in doing any of the above, some keyword-searches may be used (e.g., "action required") and some additional verifications may take place (e.g., whether the "from" is a relevant sender, or some analyses of the email's text). To verify some of these hypotheses, we performed some additional tests (on May 1st—mentioned in §III-B) by sending additional emails (from a GMail account—different from **z@gmail.com*) to 𝔾's main email account (**4@gmail.com*) with different combinations of subject and texts. We report two relevant ones below:

1) Subject: "`Urgent: Please Water the Plants Before Tomorrow`"
   Text:[7] "`Dear caretaker, \n I hope this email finds you well. I am writing to remind you of an important task that needs to be taken care of before tomorrow. \n It's crucial that the plants in the balcony receive water before tomorrow to ensure their health and well-being. As you know, they rely on regular watering to thrive, and neglecting this can lead to wilting and damage. \n Your assistance in watering the plants would be greatly appreciated. Please make sure to water them thoroughly, taking care to avoid overwatering as well. \n Thank you for your attention to this matter. If you have any questions or need assistance, please don't hesitate to reach out. \n Best regards, \n Plant united.`
   Outcome: **Notification not triggered.**

[4] Actually, we believe that it is doing a good job: the notification is very accurate! The only problem is that it can be maliciously exploited!

[5] E.g., "MSc Information Systems" (i.e., the subject of the email—see Fig. 2) never appears in the emails received by **4@gmail.com* account.

[6] Even if the "GMail notifications" are disabled, the "Google Assistant notifications" would still be shown because it is a different Android process.

[7] To generate the text, we used ChatGPT 3.5 Turbo with the prompt "write an email reminding the recipient to water the plants before tomorrow".

2) Subject: "`Water the plants - Action Required by Tomorrow`" Text: (the exact same text of the message reported above). Outcome: **Notification triggered.**

From these tests we can conclude the following:

- The triggering element is the term "Action Required by tomorrow" in the subject (we tried with "urgent" but it does not trigger the notification by Google Assistant)
- The content of the email does not appear to be very relevant (even if the name of the recipient is not specified, the notification is still triggered by Google Assistant).

However, there could be other elements that may contribute to the notification; moreover, we are also unsure about whether such notification would be triggered if the smartphone is configured to use a different language. Unfortunately, only Google knows the exact answer to all of these questions.

C. Transparency (opting-out is not straightfoward)

The issue we brought to light has its roots on the Google Assistant ecosystem and, specifically, on the *notifications* that the Google Assistant sends to users of Android smartphones. One way to nullify this problem is by **disabling the notifications** generated by the Google Assistant process. However, disabling *all* of its notifications may be impractical: some functionalities of the Google Assistant are not impacted by this issue, and some users may still want to benefit from such tools.

Given the above, we scrutinize if there is a **way to opt-out of specific functionalities** provided by Google Assistant. We found that this is not simple. As a case study, we consider how this can be done on $\mathbb{G}$'s main smartphone—a Samsung Galaxy S23 running Android 13; we performed these operations on March 18th 2024. We began by looking at the "settings" of the device, but there was no mention of the Google Assistant; even by searching for specific Apps, nothing showed up by typing "Google Assistant" in the search bar. We then decided to turn to Google itself: we performed a Google-search with the term "disable google assistant", and found a potential solution at [26]. However, the instructions in [26] did not mention anything related to the problem discussed in our paper, so we used this article only to determine how to access the menu of Google Assistant. After following these instructions, we eventually found and opened the Google Assistant menu: *we were shown 41 submenus* (shown in Fig. 8) describing the various functionalities of Google Assistant. We were overwhelmed by all these options: potential candidates were: "Accounts", "Calendar", "General", "Notes & Lists", "Notifications", "Personal Results", "Tasks", "Your Apps", "Your data in the Assistant".

We went through **all the 41 submenus of the Google Assistant**. Eventually, we found that the best fit was "Notifications": after tapping on it, we were shown four options (see Fig. 9a), from which we selected "Help with tasks" (which we found confusing that it was not put under the "Tasks" submenu). This subsubmenu began with the description "Info about your flights, bills, package deliveries & more from your Gmail or Google Account" (which substantially overlaps with the 41 aforementioned submenus.) Nevertheless, this subsubmenu includes 15 on/off options, *all enabled by default* (shown in Figs. 9b to 9d): by deselecting the "*Due date reminders*" (whose description reads "Expiring credit cards, library books & more") it is theoretically possible to opt-out from the notifications that would stem from our discovered issue. We found this whole process *impractical and convoluted*—but this is our opinion.

VI. User Study (are Android users aware of this?)

To get an understanding of how "widespread" this issue is, we carry out a user study wherein we gauge the awareness of Android users on the aspects pertaining to our underscored problem and discussed in this paper. We first describe how we carried out our study (§VI-A); then, we present the quantitative results (§VI-B) and provide a qualitative analysis (§VI-C).

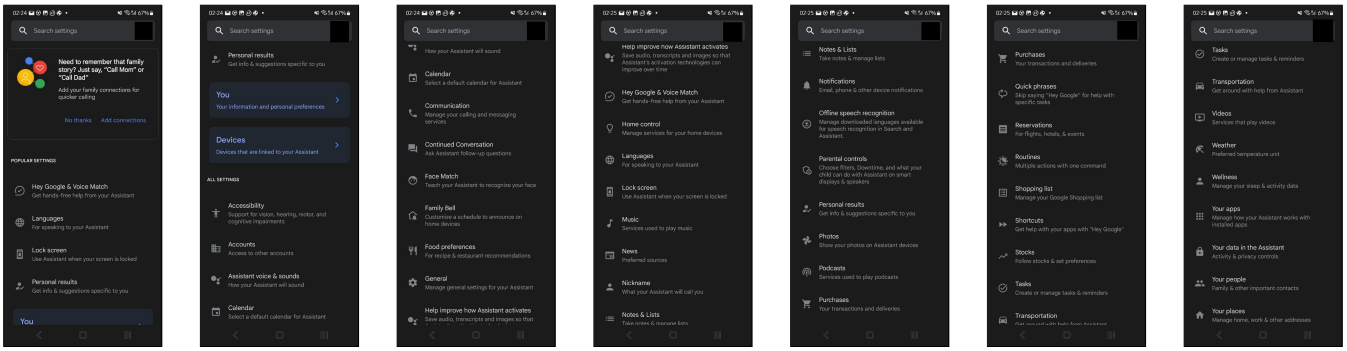
A. Study Description and Methods

Our user study consists in a survey among owners of Android devices. At a high-level, we designed an (anonymous) questionnaire wherein we ask various questions, and then distribute the questionnaire in various communities.

Questionnaire. We create the questionnaire on Google Forms, and its layout is structured in four sections consisting of mostly closed questions, which we describe below.

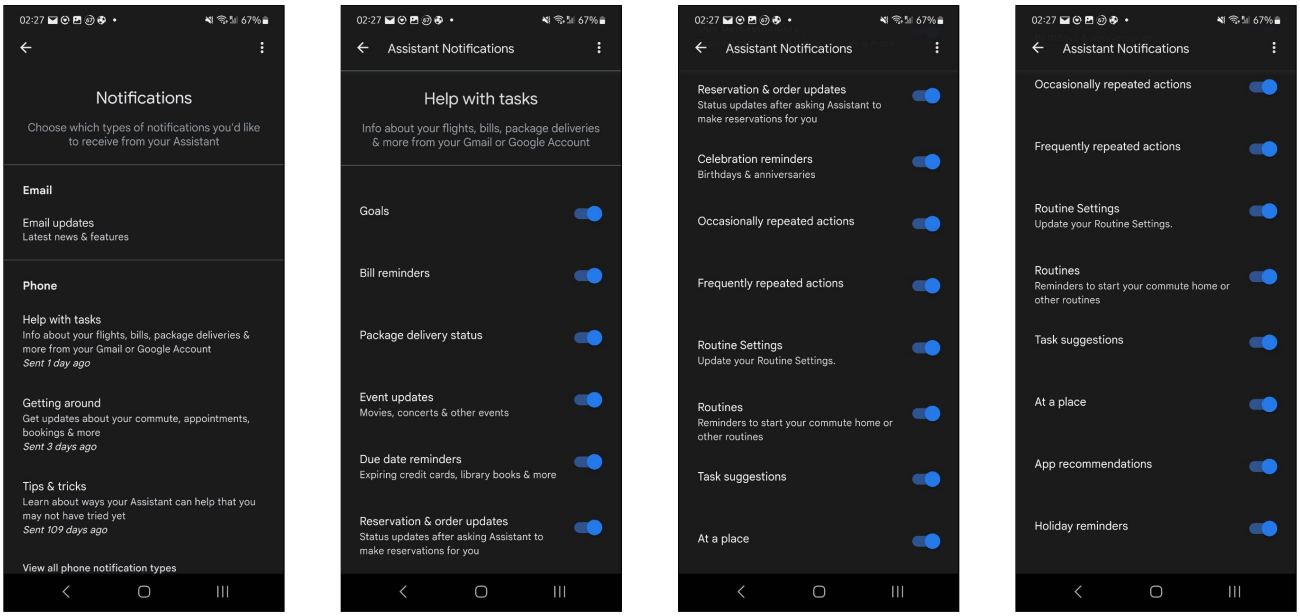
1) The participant is first informed of the purpose of our survey. We explicitly mention that the survey is reserved for people who "*own and use an Android smartphone*". We also provide our contact details, state that the questionnaire takes ≈5minutes,[8] and request the consent to participate.
2) We ask various questions about the *demographics*, such as age, country, whether the participant is tech-savvy, and whether they have an Android smartphone. If the response to this last question is "no" the survey ends; otherwise, the participant is brought to the next section.
3) We inquire about the *participant's relationship with Android.* First, we ask to provide the OS version of their (primary) Anrdoid phone—we also include an option for "do not remember". Then, we ask if the participant has an email account through which they *(i)* use Google services, and which is *(ii)* linked to their Android smartphone. Finally, we show the logo of Google Assistant, and inquire the participant if they had ever seen such a logo before.
4) The fourth section revolves around the *Google Assistant*. First, we ask if the participant knows what the Google Assistant is. Then, we inquire if the participant "recognized that the logo shown before represented the Google Assistant". Next, we ask the participant *(a)* if they know whether the Google Assistant is active (or not) on their Android smartphone; and *(b)* if they have ever checked how to disable some specific functionalities of the Google Assistant. Finally, if the participant answered "yes" to the last question of the previous section, we ask: "did you know that Google Assistant triggers a notification if you receive

[8] We carried out some pilot tests with colleagues to derive such an estimate.



(a) Page 1/7 (b) Page 2/7 (c) Page 3/7 (d) Page 4/7 (e) Page 5/7 (f) Page 6/7 (g) Page 7/7

Fig. 8: **Main Menu (of 𝔾's Samsung S23) pertaining to the "Google Assistant" options.** Overall, there are 41 "submenus" of settings, organized sparsely. To disable the functionality related to our vulnerability, one must open the "Notifications" tab (and not, e.g., the "tasks" or "account" tabs).



(a) Second Menu (b) Third Menu (1/3) (c) Third Menu (2/3) (d) Third Menu (3/3)

Fig. 9: **Follow-up menus (from 𝔾's Samsung S23) pertaining to the "Google Assistant" options.** After opening the "Notification" submenu (from Fig. 8e), we are first shown the tab in Fig. 9a. Then, after tapping on "Help with Tasks", we are shown a subsubmenu having 15 options (Figs. 9b to 9d), *all enabled by default*. To prevent exploitation of the vulnerability discussed in this paper, one could theoretically opt-out from the "Due Date Reminders".

an email with the subject 'Action required by tomorrow', even if you have never opened the email?"

We then thank the participant, and invite them not to mention the specifics of the questionnaire (to avoid biasing future responses). Finally, in an attempt to *kickstart an educational campaign* on this issue, we: *(i)* provide the links to our demonstrative video, to *(ii)* a document summarising our discovery—consisting in an early draft of the introduction of this paper; and *(iii)* invite the participant to give us their contacts if they are interested in knowing more about this issue. For scientific transparency, we provide an (anonymised) copy of our questionnaire in our repository [15].

Data Collection. We distribute our survey across various online social networks (e.g., LinkedIn, Twitter) as well as among our own personal network of contacts. Furthermore, we created two additional variants of our questionnaire in different languages (German and Italian) to allow even people who are not familiar with English to provide their input. We also kindly invited anyone who came across our survey (including participants) to share the questionnaire among their own networks to further extend the reach. This form of distribution is typically referred to as "convenience sampling" [27] (and has been also used in, e.g., [28–30]) it is appropriate for our survey given its *preliminary and exploratory nature*. We first shared our questionnaire at the end of May 2024, and kept collecting responses until the end of June 2024.

B. Survey Results (quantitative)

We received a total of 124 responses. Among these, we had to remove 12 since they pertained to users who did not own an Android device. In what follows, we quantitatively present our findings and explain why they are important.

Which Android OS do you have? This question (the first of the third section) serves to assess if our participants have devices compatible with Google Assistant (some Android devices may not support it: see [31]), and whether such devices align with those considered in our paper. Overall, 80 (71%) participants have Android 9 or higher—which align with our experiments. Intriguingly, 23 (21%) participants do not know the OS version and not even recall when they purchased the phone—indicating that they may not be fully aware of some crucial elements of the Android ecosystem.

Is your smartphone linked to an email account through which you use Google services, and is the Google Assistant active on your smartphone? These questions serve to further prove if the participants' smartphones can be affected by the problem considered in this paper. Indeed, even if someone has an Android device which supports the Google Assistant, if the Google Assistant cannot access their inbox (or is disabled in the first place) then the user would not be subject to social engineering attacks exploiting the Google Assistant. The results are enlightening. For the former question (the second of section three): 1 participant (1%) answered "I do not know"; 6 (5%) either do not use Google services, or their Android smartphone is not linked to any such email; and 105 (94%) participants answered positively—meaning that the wide majority of our sample *could* be affected by the problem we highlighted. For the later question (the third of section four): 39 (35%) are certain that the Google Assistant is disabled on their smartphone, whereas 37 (33%) think that it is enabled and 36 (32%) do not know whether Google Assistant is enabled.

What do you know about the Google Assistant? These questions serve to investigate whether our participants have any generic awareness of the main "subject" of our issue. First, with regard to the logo (last question of section three), 17 (15%) "do not recognize it" and 36 (32%) "have seen it, but do not remember what it stands for", whereas 59 (53%) "have seen it and know what it stands for". These responses are validated by the second question of section four, asking whether the participant recognized the logo as representing the Google Assistant: 67 (60%) answered "yes" and 45 (40%) answered "no". When asked about whether they know what the Google Assistant is (first question of section four), 91 (81%) answered "yes" whereas 21 (19%) answered "no".

Are you aware of this issue (and opting out)? These questions serve to explore whether the user are cognizant of the specific issues tackled in this paper. First, with regard to knowing how to disable some functionalities of the Google Assistant (fourth question of section four), and 64 (57%) "do not know" or "have never checked" how to do so, whereas 41 (37%) have checked (7 do not have Google Assistant, so they are excluded). Then, with regards to being aware that the Google Assistant triggers a notification if "Action Required" is in the email subject,[a] we found that 22 (21%) participants are aware—whereas 83 (79%) *are not aware*.

[a] Recall (see §VI-A) that this question is only for those who, according to the previous questions, can be subject to such a notification (105 in total).

C. Considerations and Explanations

We now qualitatively analyse the results of our user study. Before doing so, however, we must make some disclaimers.

Ethics and Limitations. Our institutions are aware of our research, and we carried out our user study by following established ethical guidelines [32]. Participation in our survey is voluntary and we do not offer any compensation. No harm is done to our participants. In our survey, we never ask for personally identifiable or sensitive information [33, 34]. To oblige with existing regulation, we inquire users to input a "custom string" that we can use to fulfill potential data-deletion requests we may receive afterwards. Providing the email address is not necessary and only serves to disseminate future developments of our research—which are meant to respect the right to be informed (about this subtle issue) of Android users [35]. Our user study is preliminary in nature: the worldwide population of Android users is in the billions [1], and we do not seek to generalize (which is an unfeasible goal). Indeed, our survey is exploratory (§VI-A) making our dissemination methods appropriate. The wide majority of our participants (77%) are from Europe—with the top-3 most participating countries being Italy (28%), Germany (13%) and Switzerland (13%); only 9% of our participants are from North America, whereas 13% reside in other areas of the world (and 1% preferred not to say). Hence, our results are biased towards Europe. Moreover, 35 (31%) respondents of our sample are younger than 30 years, whereas 36 (32%) have between 30–39 years, and 39 (35%) are older than 40; 2% preferred not to say. Hence, from this perspective, our population is relatively well-balanced. Finally, we did not know the opinion of any of our participants before inviting them to participate in our survey (i.e., we did not "cherry pick" communities to favor any specific outcome). To the best of our knowledge, this is the first user study with a similar design and purpose.

Interpretation. Let us try to interpret our findings. First, it is apparent that *most participants are not aware* of the subtle issue tackled in this paper; moreover, we assert that even for the few (21%) who are aware that a notification is triggered by Google Assistant if certain conditions are met, they *may not be aware that this property can be exploited for social engineering attacks*. Second, we found some inconsistencies in some of the responses (e.g., most know about Google Assistant, but they do not seem to be aware of what it truly is/does) we received—also confirmed by some personal interactions with participants who reached out to us after filling our survey. We posit that this is because people may not be cognizant of the true nature of the Google Assistant: for instance, some may believe that the Google Assistant is merely "the AI that you can talk and provide commands to" (e.g., when saying "Hey Google" out loud), whereas this is just one of the functionalities provided by the Google Assistant—which, as we showed in this paper, monitors a much larger portion of our "digital lives." Due to this consideration, we also argue that Android users may not be fully aware of the data that is analysed by the Google Assistant. Some (or, probably,

most) of our participants may know that their data is sent to Google; however, they may not know *what* data, and *how* such data is used by Google. For instance, they may not be aware that the textual content of their emails is inspected (which is something we proved in §V-A). Nonetheless, we stress that all of the above are just our educated guesses.

Further Analyses. We conclude this section by further analysing our results. Specifically, we are interested in dissecting the responses of participants who consider themselves "savvy" in information technology (IT) and of those who do not. This is instructive to discern if the issues we brought to light may affect certain groups more (or less) than others. To this end, we consider the answers to the two last questions of our questionnaire, since they are the most relevant for the sake of this paper. Let us discuss these results.

Fig. 10: **Awareness of the issue w.r.t. being (or not) savvy in IT.** The sankey shows how many participants of our user study who are IT savvy (or not) are also aware (or not) that the Google Assistant triggers a notification when a new email with "Action Required" in the subject is received. We only consider responses from participants who have their Android smartphone linked to an email account through which they use Google Services (see §VI-A).

- *Awareness of the Notification*. We report the distribution of the answers in Fig. 10. We can see that the majority of those (22) who are aware consider themselves as "IT savvy", whereas the majority of those who are not IT savvy are also not aware of this issue—which is expected. However, we also find an intriguing phenomenon: roughly half of those (83) who are not aware of this issue consider themselves as IT savvy! Such a finding shows that raising awareness that the notifications of Google Assistant can be maliciously exploited must be done throughout the whole Android userbase—and not only for those who are IT savvy!
- *Disabling the functionalities of Google Assistant*. We report the distribution of the responses in Fig. 11. We can see that, among those (41) who have checked how to disable some functionalities of Google Assistant, the wide majority also is aware of the Google Assistant and they also consider themselves as IT savvy. However, an intriguing finding is that, among those (64) who have never checked (or do not know) how to disable some functionalities of the Google Assistant, the majority claims to know the Google Assistant. Such a finding may confirm our previous hypotheses: Android users may "know" the Google Assistant at a high-level, but they are not aware of its ins-and-outs—suggesting that there may be an overall lack of transparency on this crucial process of the Android ecosystem (see also our analysis in §V-C).

In light of the above, we hence conjecture (we cannot make statistically significant conclusions due to our limited sample size) that the Android userbase may overlook pivotal elements of the ecosystem that empowers their smartphones.

VII. Discussion (what can be done to fix this?)

Our paper underscored a blind spot in the Android ecosystem, which cannot only be maliciously exploited, but which also does not appear to be well-understood by the very owners of Android smartphones. Hence, to address this issue, we first propose two "patches" that can mitigate this subtle problem (§VII-A). Then, we outline how our findings can be leveraged by future work in related domains (§VII-B).

A. Mitigation (joint effort is required)

As we discussed (§V-C), opting-out from the "Due date reminders" allows Android users to be protected against our identified security issue. However, such a solution *(i)* is not straightfoward: users may not even know how to do so (§VI-C); and *(ii)* presents tradeoffs: maybe some users want such reminders. Hence, to fix this problem, we propose alternative ways which we coalesce in two "patches."

The First Patch entails *making users aware that the Google Assistant checks their inboxes* to trigger (some) notifications. In this way, end-users cannot only "responsibly" determine whether to disable[a] such functionality (e.g., for privacy reasons, or to protect themselves against attacks leveraging our discovery), but would also be "trained" to be more cautious when tapping any given notification—which may lure them to a phishing email.

[a]This process can also be made more straightforward, and potentially enhanced by allowing users to opt-out of reminders based on unread emails.

The Second Patch entails *improving the analytical abilities of the Google Assistant.* As we demonstrated, the email that triggered the notification discussed in this paper had plenty of "suspicious" indicators (e.g., mismatched name/address, recent email, weird link). Hence, by leveraging various sources (e.g., some cross-checks in the "from" field, as well as the age of a given sender email address), the Google Assistant can be turned into a form of "phishing detector" and potentially warn the user that they may be subject to a social engineering attack.[a]

[a]Besides, such notifications need not be "instantaneous": even if they are sent a few minutes after reception of an email (to improve accuracy with more analyses [36]), users would still find them useful.

Nevertheless, our proposed mitigations require a cooperative effort from the security community. As we showed (§IV), the Google Team is not going to take action. We embrace the recommendations of ethical disclosure, underlining the "right to be informed" of end-users [32, 35]. Hence, we advocate our community to raise awareness about this issue (unfortunately, our preliminary user study is only a first step, and more effort is required by our community). In this way, the developers of Google may have more incentives to take action to *(i)* simplify the opt-out process, *(ii)* devise educational campaigns for Android users, and *(iii)* improve the Google Assistant.

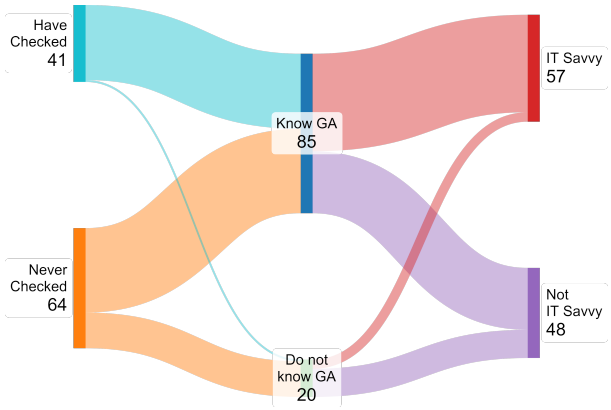


Fig. 11: **Distribution of responses for those who "have checked" (or not) how to disable some functionalities of the Google Assistant (GA).** For those who "have not checked", we group the responses of those who explicitly mentioned of not having checked together with those who "do not know." We exclude the 7 participants who do not use GA on their smartphone.

B. Implications for future work on Socio-technical Security

The issue we brought to light touches various aspects of socio-technical security (and privacy). We summarize five areas of related work that can be inspired by our discovery and build upon our findings, suggesting avenues for future work.

- **Human Factor in Cybersecurity.** At its core, our identified issue must be exploited via social-engineering attacks, i.e., by tricking humans into releasing credentials, or clicking malicious links. Hence, our discovery has strong connections with all works that explore the role of the human factor in cybersecurity [37] and, in particular, in phishing attacks [28, 38]. Intriguingly, the problem we underscored entails leveraging a *benign functionality for malicious purposes* (which is similar to the scenarios envisioned in [39]).
- **Security of AI.** It is well-known that the Google Assistant uses AI [5, 6]. Even though we cannot be certain that the notifications for "Due date reminders" are generated by means of some AI model, there is reason to believe that this is the case (after all, analysing text is one of the primary uses of AI). Hence, our findings are strongly related to studies on attacks against AI-based systems (e.g., [40]).
- **Privacy of AI.** Our analyses revealed that "some AI" may be analysing private data (i.e., emails). Of course, some Android users may be willing to have an AI read their emails and generate notifications to improve their overall quality of life; yet, others may not be of the same opinion. The crux, however, is that all users must be adequately informed of these operations [41]. This is also necessary in light of upcoming regulation [29, 42]. Hence, our paper also seeks to build a bridge with the privacy domain—and, specifically, the one focusing on Android [43]. As we showed (§V-C), the numerous functionalities of Google Assistant are not easy to understand, and most options are enabled by default. Remarkably, even Google is interested in providing more privacy-friendly AI technologies [44].
- **Mobile (and Web) Security.** Our discovery pertains to the security of the Android (and the Web) ecosystem, hence many connections can be made with these well-studied research domains (e.g., [45–52]). Interestingly, however, we found that the specific topic tackled by our paper ("phishing on Android") is relatively under-investigated.[9] The most notable work is [12]: here, Ruggia et al. envision a "reversed" scenario as the one considered in our paper. Specifically, Ruggia et al. [12] seek to exploit the notification mechanism of Android to "notify" a phisher that their victim has opened a certain App, and then use such information to deliver a well-crafted "phishing hook" to the victim.
- **Offensive AI.** Recent works have raised the attention on the fact that AI methods can be leveraged by evildoers to convey cyberattacks [58, 59]. In a sense, the security issue discussed in our paper can fall in this category: first, because the Google Assistant uses AI, and exploiting its notifications is a way to "maliciously exploit" an AI (at the expense of Android users); second, because it is possible to use AI to generate the text of the email that will convey the phishing attack [60]. Indeed, in some of our analyses (e.g., the checks in §V-B, or for our demonstrative video [15]), we also did use ChatGPT to generate the email. Hence, our paper is a case study of how AI can act as criminal co-conspirator.

Intriguing follow-up of our work can entail *carrying out additional user studies* [41, 61], which expand our preliminary study (§VI) by, e.g., further investigating how much Android users know about the Google Assistant ecosystem, or educating end-users on this subtle threat. Future work can also explore the *development of technical countermeasures* [62–64] that, e.g., improve the capabilities of existing systems to detect the malicious email triggering the Android notification. Both of these avenues align with our "patches" (§VII-A).

VIII. Conclusions

We discovered that a benign functionality of the Google Assistant ecosystem can be maliciously exploited. In this way, evildoers can send phishing emails soliciting their targets (i.e., Android users) to "take action soon". Such emails will be read by the Google Assistant, which will generate a notification that urges the end-user that "a task is soon due." The user will hence be more likely to fall victim to social engineering attacks—driven by such a sense of urgency.

Despite disclosing this issue to Google—who acknowledged this possibility—they responded that a fix will not be implemented. We verified that this problem is still exploitable at the time of writing this paper (May 2024). Moreover, we showed that Android users are hardly aware of the functionalities of the Google Assistant. We hence endorse the security community to raise awareness on this issue, so that Android users are informed of this problem, but also to encourage the development of "patches" by Google that make Android more secure.

[9] **Systematic Literature Review.** To validate this claim, we analysed the proceedings from 2014 to 2023 of 11 top-venues related to Security, Human Factors and the Web: WWW, S&P (and EuroS&P), CCS, USENIX SEC, NDSS, AsiaCCS, ACSAC. IMC, WSDM, CHI. We searched for full papers (excluding, e.g., workshops) having "phish" in the title and found 56 papers. Then, we inspected their text (excluding references), searching for occurrences of "Android", and found 9 papers (out of 56). However, 5 papers (i.e., [10, 53–56]) only mention "Android" once/twice (typically out of context). Papers with more hits are [9, 11, 12, 57], which indeed focus on phishing in the Android ecosystem—representing only 7% of the papers we analysed.

⚠ **Why is this a Problem?** To conclude, we highlight three points that aggravate the "risk" of our envisioned attack:

- the notification is *independent* of the Calendar (Footnote 1) or GMail (Footnote 6) Android applications;
- the *only requirement* is that users have Google Assistant enabled and linked to an email account related to Google;
- the notification is triggered by *an email that is not (nor needs to be) read* by the user (see Fig. 15a).

Put simply, users are suddenly reminded to "do something *now*", but they are oblivious of the reason. This "urgent confusion" may induce users to overlook that they are targeted by a phishing attack. Awareness of such a risk is paramount to protect Android users against social engineering.

ACKNOWLEDGEMENT. The authors would like to thank: Aurore Fass and Mia Gandini for extensive feedback that improved this research immensely; the participants of our user survey for their contributions; and the Hilti Group for funding.

POST-ACCEPTANCE ADDENDUM (BUGFIX)

(This section has been added after acceptance of this paper)

While preparing our presentation for eCrime (Sept. 2024), we found that we were unable to reproduce the issue discussed in the paper. We tried all the emails that worked before, and also tried to make changes in the subject and text; however, all these attempts were not successful: the email was received, but no "Google Assistant Notification" was generated.

We investigated, and found that a potential explanation may be related to a recent change in the Google Assistant ecosystem: in August 14th, Google announced that "Gemini," their new LLM, is now the Google Assistant on Android smartphones [65]. Such a change may have also led to patching some behavior of the "old" Google Assistant.

Nevertheless, we tried asking Gemini to reproduce the behavior (see Figs. 12). Surprisingly, Gemini refused to do so, saying that "it isn't secure" – which is something we agree with. However, we find that this response somewhat contradicts what transpired during our exchange with Google in January 2024: recall (§IV) that our bug report was closed with the motivation that the issue was "intended behavior" (this begs the question: was it really intended that the Google Assistant could access our personal data and automatically decide to send us reminders about tasks we were oblivious of?). Nonetheless, we are glad that such a functionality does not appear to be operational (or, at least, enabled by default) on current Android phones. However, the change to the Google Assistant should inspire future work to investigate the security of Gemini: as our user survey highlighted, users are not aware of the detailed capabilities of these AI assistants.

REFERENCES

[1] BankMyCell. (2024) How many android users are there? global and us statistics. https://web.archive.org/web/20240205082953/https://www.bankmycell.com/blog/how-many-android-users-are-there.

[2] StatCounter. (2024) Mobile operating system market share worldwide. https://web.archive.org/web/20240318202847/https://gs.statcounter.com/os-market-share/mobile/worldwide.

[3] Statista. (2024) Leading tech companies worldwide 2024, by market capitalization. https://web.archive.org/web/20240130130259/https://www.statista.com/statistics/1350976/leading-tech-companies-worldwide-by-market-cap/.

[4] Google. (2020) A more helpful Google Assistant for your every day. https://web.archive.org/web/20201127165358/https://www.blog.google/products/assistant/ces-2020-google-assistant/.

[5] CultOfMac. (2016) The future is AI, and Google just showed Apple how it's done. https://web.archive.org/web/20201108162911/https://cultofmac.com/447898/google-home-google-assistant-siri-ai.

[6] Google. (2023) Assistant with Bard: A step toward a more personal assistant. https://web.archive.org/web/20240222161116/https://blog.google/products/assistant/google-assistant-bard-generative-ai.

[7] Droid-life. (2023) Google assistant reminders go live in google tasks: How to switch. https://web.archive.org/web/2/https://www.droid-life.com/2023/06/01/google-assistant-reminders-go-live-in-google-tasks-how-to-switch/.

[8] T. N. News. (2021) Google assistant now provides ai-powered mental health support to arabic speakers. https://web.archive.org/web/2/https://www.thenationalnews.com/lifestyle/wellbeing/google-assistant-now-provides-ai-powered-mental-health-support-to-arabic-speakers-1.1192581.

[9] S. Aonzo, A. Merlo, G. Tavella, and Y. Fratantonio, "Phishing attacks on modern android," in *ACM CCS*, 2018.

[10] K. Tian, S. T. Jan, H. Hu, D. Yao, and G. Wang, "Needle in a haystack: Tracking down elite phishing domains in the wild," in *IMC*, 2018.

[11] C. Marforio, R. Jayaram Masti, C. Soriente, K. Kostiainen, and S. Čapkun, "Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications," in *ACM CHI*, 2016.

[12] A. Ruggia, A. Possemato, A. Merlo, D. Nisi, and S. Aonzo, "Android, notify me when it is time to go phishing," in *IEEE EuroS&P*, 2023.

[13] "State of the phish," https://www.proofpoint.com/it/resources/threat-reports/state-of-phish, ProofPoint, Tech. Rep., 2023.

[14] E. J. Williams, J. Hinds, and A. N. Joinson, "Exploring susceptibility to phishing in the workplace," *Int. J. Human-Computer Studies*, 2018.

[15] "Our repo," https://github.com/hihey54/ecrime24/.

[16] "Gophish," https://getgophish.com/, 2024.

[17] "Zphisher," https://github.com/htr-tech/zphisher, 2024.

[18] P. Burda, A. M. Altawekji, L. Allodi, and N. Zannone, "The peculiar case of tailored phishing against smes: Detection and collective defense-mechanisms at a small it company," in *IEEE EuroS&P Workshops*, 2023.

[19] M. Lokesh, A. K. Devi, U. D. Chowdary, P. D. Lakshmi, and G. R. K. Rao, "Data redundancy, data phishing, and data cloud backup," in *IEEE ICECCT*, 2023.

[20] "Android Distribution Chart," https://web.archive.org/web/20240622231458/https://www.composables.com/tools/distribution-chart, 2024.

[21] "Google Issue Tracker," https://issuetracker.google.com.

[22] "Google individual contributor license agreement," https://cla.developers.google.com/about/google-individual, 2024.

[23] "Attacks facilitating phishing or social engineering," https://bughunters.google.com/learn/invalid-reports/invalid-attack-scenarios/6325772798918656/attacks-facilitating-phishing-or-social-engineering, 2024.

[24] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM CSUR*, 2023.

[25] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, 2023.

[26] AndroidPolice. (2024) https://web.archive.org/web/20240214222415/https://www.androidpolice.com/how-to-disable-google-assistant.

[27] P. Sedgwick, "Convenience sampling," *British Medical Journal*, 2013.

[28] 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 eCrime*, 2023.

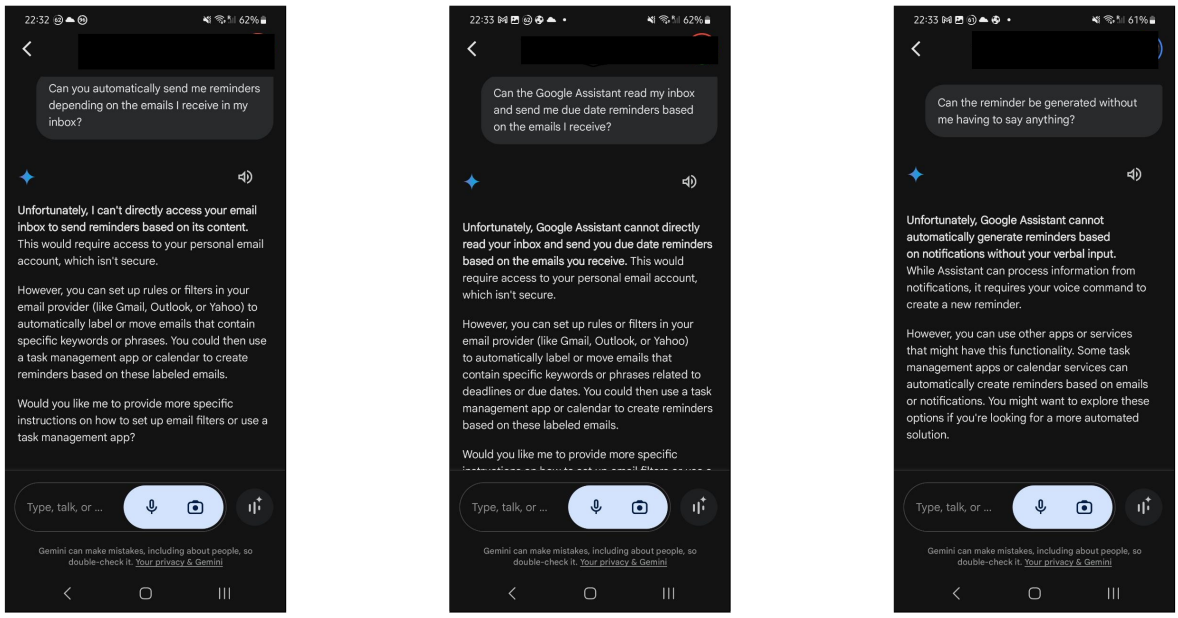
[29] F. Koh, K. Grosse, and G. Apruzzese, "Voices from the frontline: Revealing the ai practitioners' viewpoint on the european ai act," in *HICSS*, 2024.

[30] Y. Acar, C. Stransky, D. Wermke, M. L. Mazurek, and S. Fahl, "Security developer studies with {GitHub} users: Exploring a convenience sample," in *SOUPS*, 2017.

[31] Google. (2024) https://web.archive.org/web/20240622051941/https://support.google.com/assistant/answer/7172657?hl=en&co=GENIE.Platform%3DAndroid.

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

[33] E. Commission, "Sensitive data," https://ec.europa.eu/info/law/law-



(a) First inquiry. (b) Second inquiry. (c) Third inquiry.

Fig. 12: **Conversation with Gemini.** We asked Gemini, which replaced the Google Assistant in Aug. 2024 [65], if there were ways to reproduce the behavior that enabled the vulnerability. According to Gemini, this is not possible because "it isn't secure", but it's possible if the user explicitly asks for it (Fig. 12c)

topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en.

[34] U. D. of the Treasury, "Sensitive personal data," https://home.treasury.gov/taxonomy/term/7651.

[35] T. Kohno, Y. Acar, and W. Loh, "Ethical frameworks and computer security trolley problems: Foundations for conversations," in *USENIX Sec.*, 2023.

[36] J. Lee, Z. Xin, M. P. S. Ng, K. Sabharwal, G. Apruzzese, and D. M. Divakaran, "Attacking logo-based phishing website detectors with adversarial perturbations," in *ESORICS*, 2023.

[37] B. Dupont and T. Holt, "The human factor of cybercrime," *Social Science Computer Review*, 2022.

[38] S. Baki and R. M. Verma, "Sixteen years of phishing user studies: What have we learned?" *IEEE TDSC*, 2023.

[39] Y. T. Chua, S. Parkin, M. Edwards, D. Oliveira, S. Schiffner, G. Tyson, and A. Hutchings, "Identifying unintended harms of cybersecurity countermeasures," in *APWG eCrime*, 2019.

[40] L. Pajola and M. Conti, "Fall of Giants: How popular text-based MLaaS fall against a simple evasion attack," in *IEEE EuroS&P*, 2021.

[41] R. Tawer, M. Mehrnezhad, and C. Morisset, "I feel spied on and i don't have any control over my data": User privacy perception, preferences and trade-offs in university smart buildings," *STAST*, 2022.

[42] N. Leesakul and C. Morisset, "Position paper: The role of law in achieving privacy and security measures in smart buildings from the gdpr context," in *IEEE EuroS&PW (STAST)*, 2023.

[43] X. Liu, J. Liu, S. Zhu, W. Wang, and X. Zhang, "Privacy risk analysis and mitigation of analytics libraries in the android ecosystem," *IEEE TMC*, 2019.

[44] ArsTechnica. (2024) Google says chrome's new real-time url scanner won't invade your privacy. https://web.archive.org/web/2/https://arstechnica.com/gadgets/2024/03/google-says-chromes-new-real-time-url-scanner-wont-invade-your-privacy/.

[45] S. Karthick and S. Binu, "Android security issues and solutions," in *IEEE ICIMIA*, 2017.

[46] J. Qiu, J. Zhang, W. Luo, L. Pan, S. Nepal, and Y. Xiang, "A survey of android malware detection with deep neural models," *CSUR*, 2020.

[47] J. Senanayake, H. Kalutarage, M. O. Al-Kadri, A. Petrovski, and L. Piras, "Android source code vulnerability detection: a systematic literature review," *ACM CSUR*, 2023.

[48] D. R. Thomas, A. R. Beresford, and A. Rice, "Security metrics for the android ecosystem," in *Proc. ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, 2015.

[49] M. Linares-Vásquez, G. Bavota, and C. Escobar-Velásquez, "An empirical study on android-related vulnerabilities," in *IEEE/ACM MSR*, 2017.

[50] S. Hsu, M. Tran, and A. Fass, "What is in the chrome web store?" in *ACM AsiaCCS*, 2024.

[51] P. Zhang, Z. Sun, S. Kyung, H. W. Behrens, Z. L. Basque, H. Cho, A. Oest, R. Wang, T. Bao, Y. Shoshitaishvili *et al.*, "I'm SPARTACUS, No, I'm SPARTACUS: Proactively Protecting Users from Phishing by Intentionally Triggering Cloaking Behavior," in *Proc. ACSAC*, 2022.

[52] Á. Feal, P. Vallina, J. Gamba, S. Pastrana, A. Nappa, O. Hohlfeld, N. Vallina-Rodriguez, and J. Tapiador, "Blocklist babel: On the transparency and dynamics of open source blocklisting," *IEEE TNSM*, 2021.

[53] E. Ulqinaku, H. Assal, A. Abdou, S. Chiasson, and S. Capkun, "Is real-time phishing eliminated with FIDO? social engineering downgrade attacks against FIDO protocols," in *USENIX Sec.*, 2021.

[54] P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang *et al.*, "Crawlphish: Large-scale analysis of client-side cloaking techniques in phishing," in *IEEE S&P*, 2021.

[55] F. Quinkert, M. Degeling, J. Blythe, and T. Holz, "Be the phisher–understanding users' perception of malicious domains," in *ACM AsiaCCS*, 2020.

[56] M. Liu, Y. Zhang, B. Liu, Z. Li, H. Duan, and D. Sun, "Detecting and characterizing sms spearphishing attacks," in *ACSAC*, 2021.

[57] G. S. Tuncay, J. Qian, and C. A. Gunter, "See no evil: phishing for permissions with false transparency," in *USENIX Sec.*, 2020.

[58] M. M. Yamin, M. Ullah, H. Ullah, and B. Katt, "Weaponized ai for cyber attacks," *Journal of Information Security and Applications*, 2021.

[59] Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, and D. Gelei et al., "The threat of offensive ai to organizations," *Computers & Security*, 2023.

[60] T. Langford and B. Payne, "Phishing faster: Implementing chatgpt into phishing campaigns," in *Proc. Fut. Tech. Conf.*, 2023.

[61] J.-W. Bullee and M. Junger, "How effective are social engineering interventions? a meta-analysis," *Information & Computer Security*, 2020.

[62] C. Beckmann, B. Berens, N. Kühl, P. Mayer, M. Mossano, and M. Volkamer, "Design and evaluation of an anti-phishing artifact based on useful transparency," in *STAST*, 2022.

[63] L. Allodi, T. Chotza, E. Panina, and N. Zannone, "The need for new antiphishing measures against spear-phishing attacks," *IEEE Security & Privacy*, 2020.

[64] E. Badawi, G.-V. Jourdan, and I.-V. Onut, "Web scams detection system," in *Int. Symp. Found. Pract. Secur.*, 2023.

[65] Google. (2024) Gemini makes your mobile device a powerful AI assistant. https://web.archive.org/web/2/https://blog.google/products/gemini/made-by-google-gemini-ai-updates/.

APPENDIX

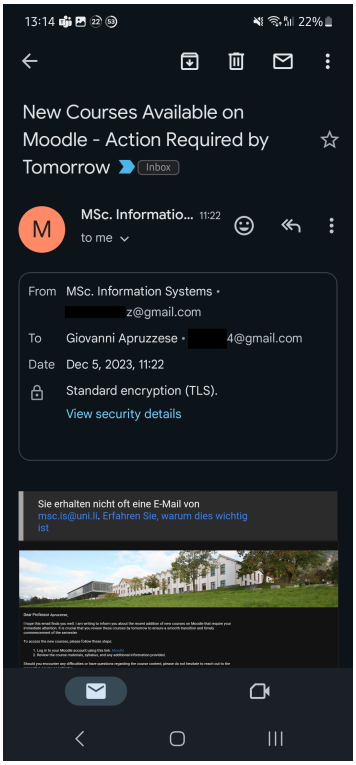
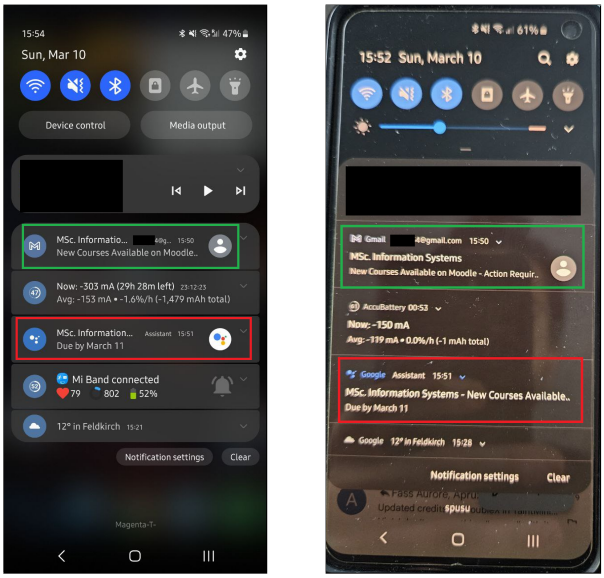
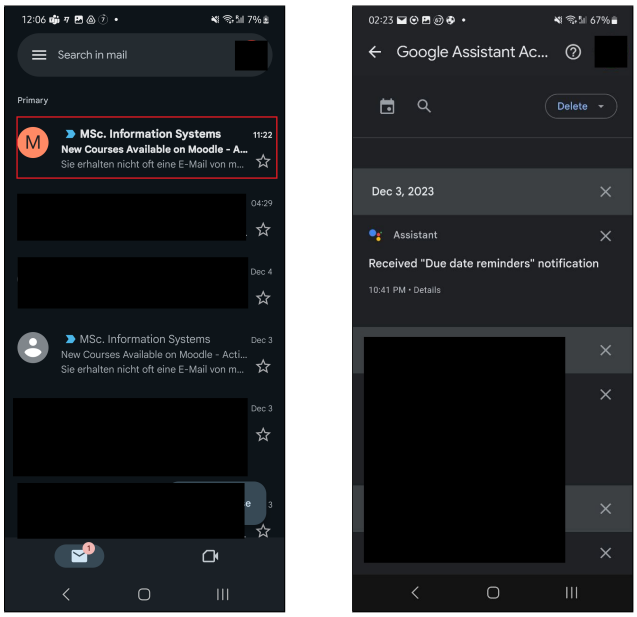


Fig. 13: **By "tapping" the notification, users are brought to the email.** When 𝔾 tapped the notification generated by the Google AI Assistant, 𝔾 was led to the inbox and the email was automatically opened.



(a) **Samsung S23.** (Android 13). (b) **Samsung S10e.** (Android 9).

Fig. 14: **Second Validation (March 10th, 2024).** We repeat the experiment in §III-A three months later (after having reached out to Google). The results (further confirmed by additional tests in May 2024) confirm that this potential vulnerability had not been fixed yet (despite Google having been informed).



(a) **The "email inbox"** of 𝔾 (**4@gmail.com*) in the morning of December 5th. The red box denotes the reception of the first email (Fig. 2).

(b) **Notification history** of 𝔾: a notification was also sent on December 3rd, 2023—related to the very first email sent by 𝕄 (see §II-B)

Fig. 15: **Additional Validation Screenshots** from 𝔾's Samsung S23.

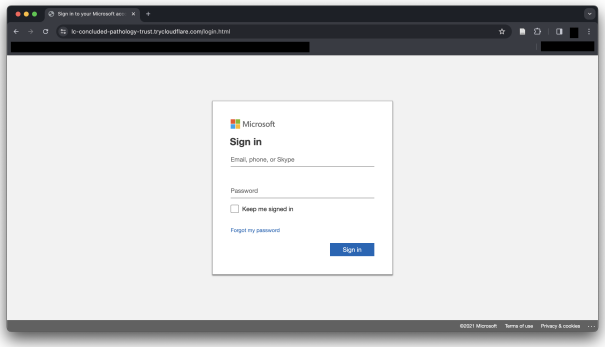


Fig. 16: **The webpage pointed to by the link in the email.** Clicking on the "Moodle" link (shown in the email in Fig. 2) leads to this webpage—crafted with Zphisher, which also logs any visit (see Fig. 5). The webpage is hosted on a Raspberry Pi4 owned and managed by 𝕄.