Barcelona – November 16th, 2023

APWG Symposium on Electronic Crime Research

Click to add title

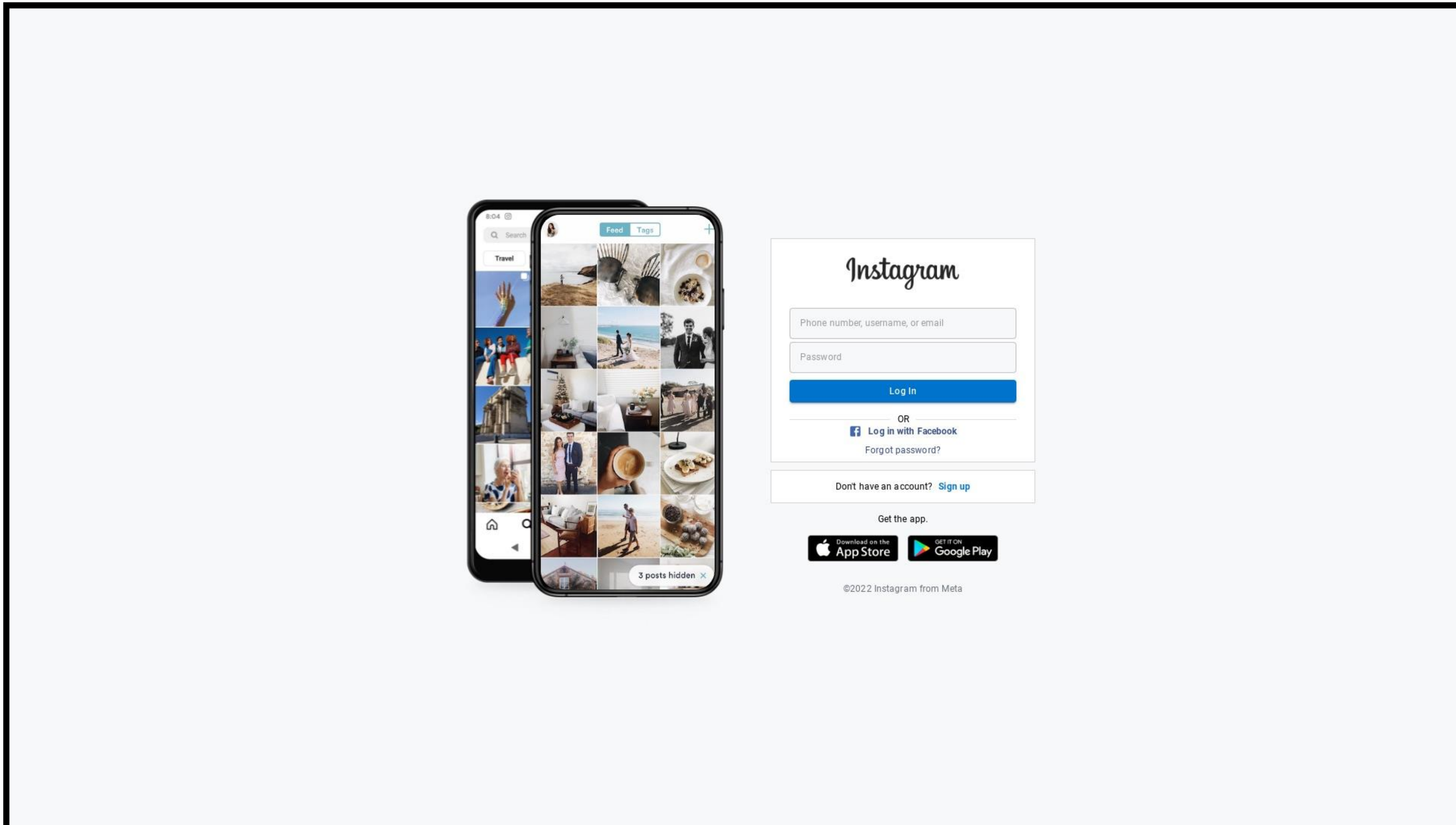
Ajka Draganovic, Savino Dambra, Xavier Aldana Iouit,
Kevin Roundy, Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

Avast

NortonLifeLock

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li



Barcelona – November 16th, 2023

APWG Symposium on Electronic Crime Research

“Do Users fall for Real Adversarial Phishing?” Investigating the Human response to evasive Webpages

Ajka Draganovic, Savino Dambra, Xavier Aldana Iouit,
Kevin Roundy, Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

Avast

NortonLifeLock

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(Phishing 101)

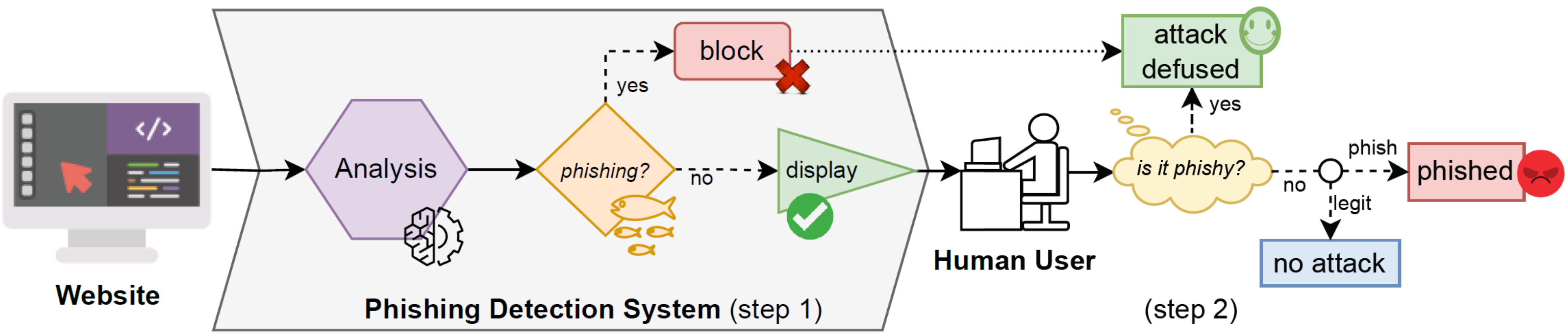


Fig. 1: Scenario: phishing detection is a two-step decision process.

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(Phishing 101)

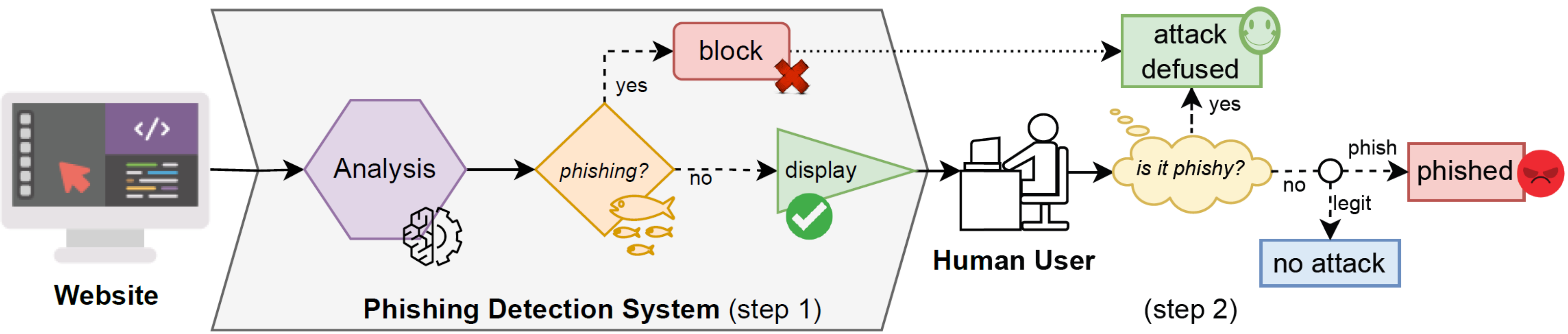


Fig. 1: Scenario: phishing detection is a two-step decision process.

We focus on Phishing Detection Systems powered by <u>Machine Learning</u>

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: Technical papers…

Typical workflow of an “adversarial machine learning” paper:

1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)

Attack

Self-developed ML model
(trained on synthetic ‘benchmark’)

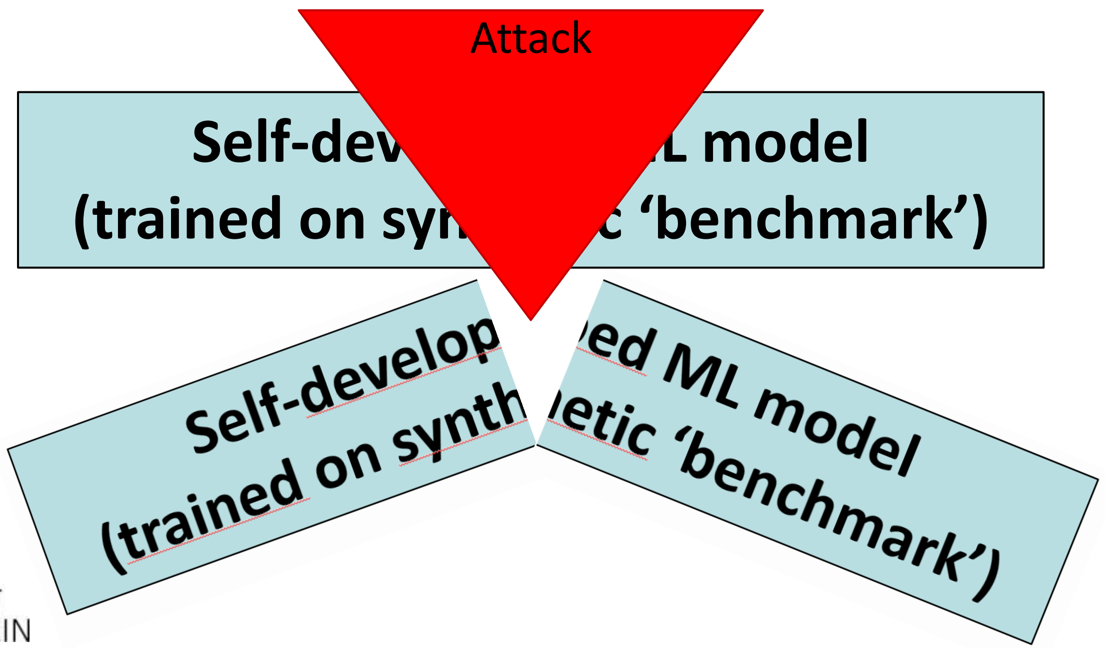
UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: Technical papers…

Typical workflow of an “adversarial machine learning” paper:

1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)
3. Show that the attack “breaks” the ML model



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: Technical papers…

Typical workflow of an "adversarial machine learning" paper:

1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)
3. Show that the attack "breaks" the ML model

What about real ML systems?

- Evading *real* ML systems is not simple [10] (and few works do this)

UNIVERSITÄT LIECHTENSTEIN

[10] 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 SaTML, 2023.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: Technical papers…

Typical workflow of an “adversarial machine learning” paper:

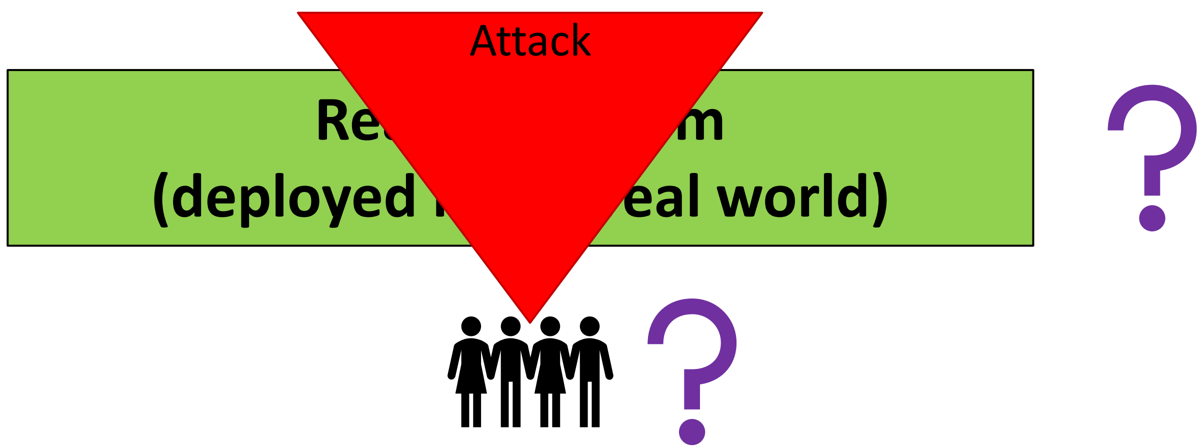
1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)
3. Show that the attack “breaks” the ML model

What about real ML systems?

- Evading *real* ML systems is not simple [10] (and few works do this)

…and are humans tricked as well?

- In some settings (e.g., phishing), humans *see* the “adversarial example”



UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: …and user studies

Typical workflow of a user study on “phishing assessment”:

1. Craft/collect phishing samples
2. Create a questionnaire and ask users to identify phishing samples
3. Draw conclusions

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: …and user studies

Typical workflow of a user study on “phishing assessment”:

1. Craft/collect phishing samples
2. Create a questionnaire and ask users to identify phishing samples
3. Draw conclusions

What about real (ML-based) phishing detectors?

- Maybe the samples would be trivially blocked by the detector

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Gap: …and user studies

Typical workflow of a user study on “phishing assessment”:

1. Craft/collect phishing samples
2. Create a questionnaire and ask users to identify phishing samples
3. Draw conclusions

What about real (ML-based) phishing detectors?

- Maybe the samples would be trivially blocked by the detector

…and what about priming?

- Users are more suspicious when they are aware of being “tested” for phishing

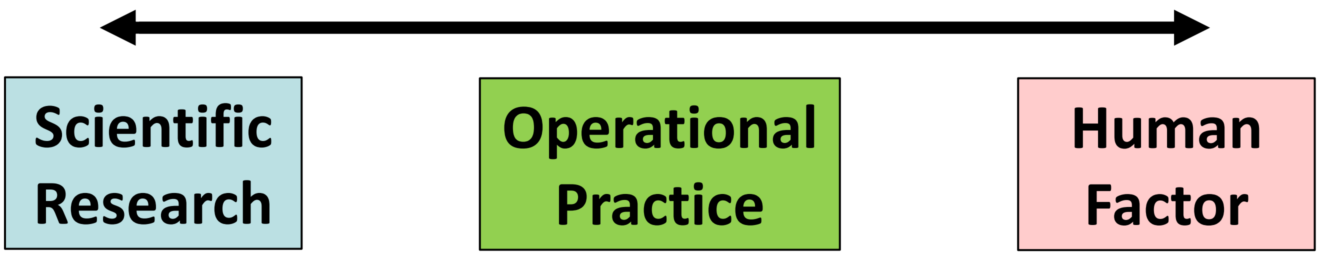
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do

We try to align

- **Research** in ML security, with
- **Operational** ML security and with
- The **human factor** in ML security



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do

We try to align

- **Research** in ML security, with
- **Operational** ML security and with
- The **human factor** in ML security

Scientific Research ⟷ **Operational Practice** ⟷ **Human Factor**

We do this by focusing on Phishing Website Detection. We consider an

- *operational ML system* (deployed in real world), which has been
- bypassed by "adversarial webpages" (crafted by *real attackers*), and
- scrutinize whether humans are *also* deceived by such evasive webpages

Nobody did this before (ttbook)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How did we do it? (1)

1. We reach out to a well-known security company ("Sigma")
2. We ask Sigma to provide us with phishing webpages that evaded their operational Phishing Detection System (reliant on deep learning)

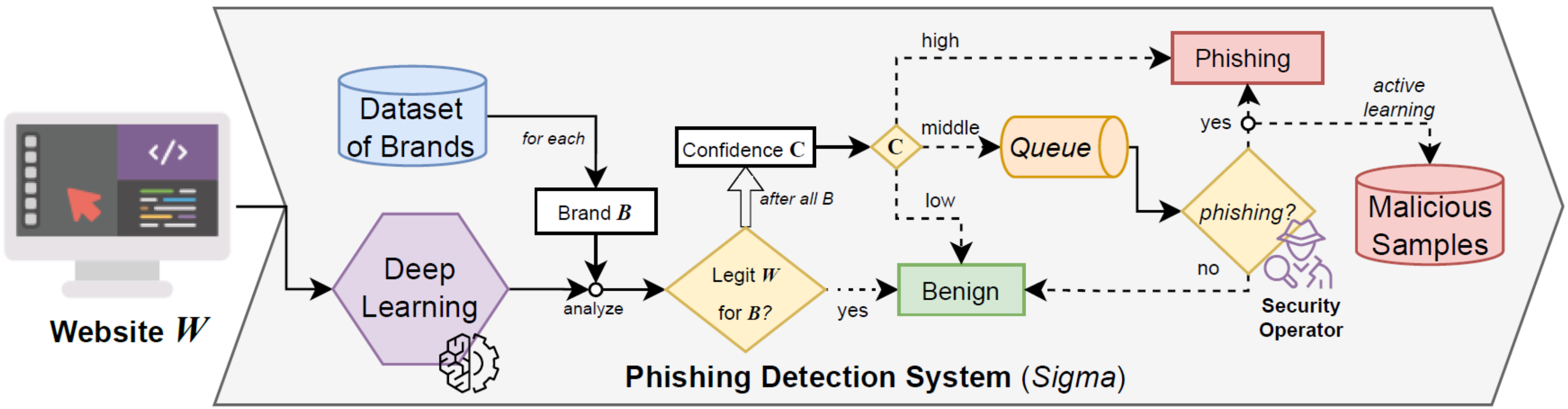


Fig. 2: The architecture of the PDS deployed by *Sigma*, used as basis for the phishing examples to include in our user-study.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How did we do it? (2)

3. We select a set of 18 “adversarial” phishing webpages (mimicking brands popular in the EU)
4. We add 2 “legitimate” webpages (as a form of control)
5. We use the screenshots of these 20 webpages to carry out a user study

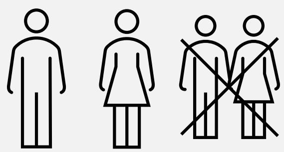
TABLE III: Sequence of screenshots in our questionnaire, and their difficulty level. The number points to the image (hosted in our repo).

#	Brand	Difficulty	Comment
1	Instagram	*Hard*	Resembles the legitimate login page, with the sole distinction being the footer’s style.
2	Facebook	*Moderate*	Appears similar to the authentic version; however, suspicion may arise due to the multiple profiles that have recently logged in from the same device (specifically, six different profiles).
3	Facebook	*Hard*	Closely resembles the original, with the sole exception of a missing footer.
4	Instagram	*Hard*	Extremely challenging to distinguish, as it perfectly mirrors the original.
5	PayPal	*Hard*	Resembles the authentic site very closely.
6	Google	*Hard*	Resembles the authentic site very closely.
7	Amazon	*Hard*	Resembles the authentic site very closely.
8	Airbnb	—	It is the legitimate website.
9	Zalando	—	It is the legitimate website.
10	Netflix	*Moderate*	The website’s header and logo may induce suspicion due to their uncharacteristic design.
11	Yahoo	*Hard*	Resembles the authentic site very closely.
12	Yahoo	*Hard*	Resembles the authentic site very closely.
13	Netflix	*Easy*	The font style noticeably deviates from the one typically used.
14	Uber	*Easy*	The appearance of Uber’s sign-in page notably diverges from the expected layout.
15	PayPal	*Moderate*	The background color of the input fields clashes with the overall design aesthetic of the website.
16	Uber	*Easy*	The appearance suggests it might be an outdated version of Uber.
17	LinkedIn	*Easy*	The font style significantly deviates from what one would expect on a professional website, disrupting its overall look and feel.
18	Netflix	*Very easy*	No resemblance to the original sign-up page, with a starkly contrasting and distinctive styling.
19	Twitter	*Moderate*	It gives the impression of being an older version of Twitter, which could still potentially elicit trust from unfamiliar users.
20	Amazon	*Moderate*	While it bears a striking resemblance, participants might grow suspicious due to the button on the page appearing incongruous with the overall design.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

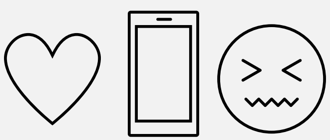
How did we do it? (3)

6. We advertise the questionnaire on popular social media for 3 weeks
7. **We do not prime the users (!)**
8. We received 126 responses



70 ***55*** ***1***

IT expertise



75 ***48*** ***3***

Age

<16 → 3

16-24 years → 44

25-34 years → 57

35-44 years → 12

45-54 years → 4

55-64 years → 6

Country

78 ***19*** ***12***

Slovenia, Italy, Bosnia, Liechtenstein, Finland …

Education

Basic → 11

High School → 45

Bachelor's → 41

Master's → 27

PhD → 2

1. Screenshot - Please rate how much you agree with the following statement: *

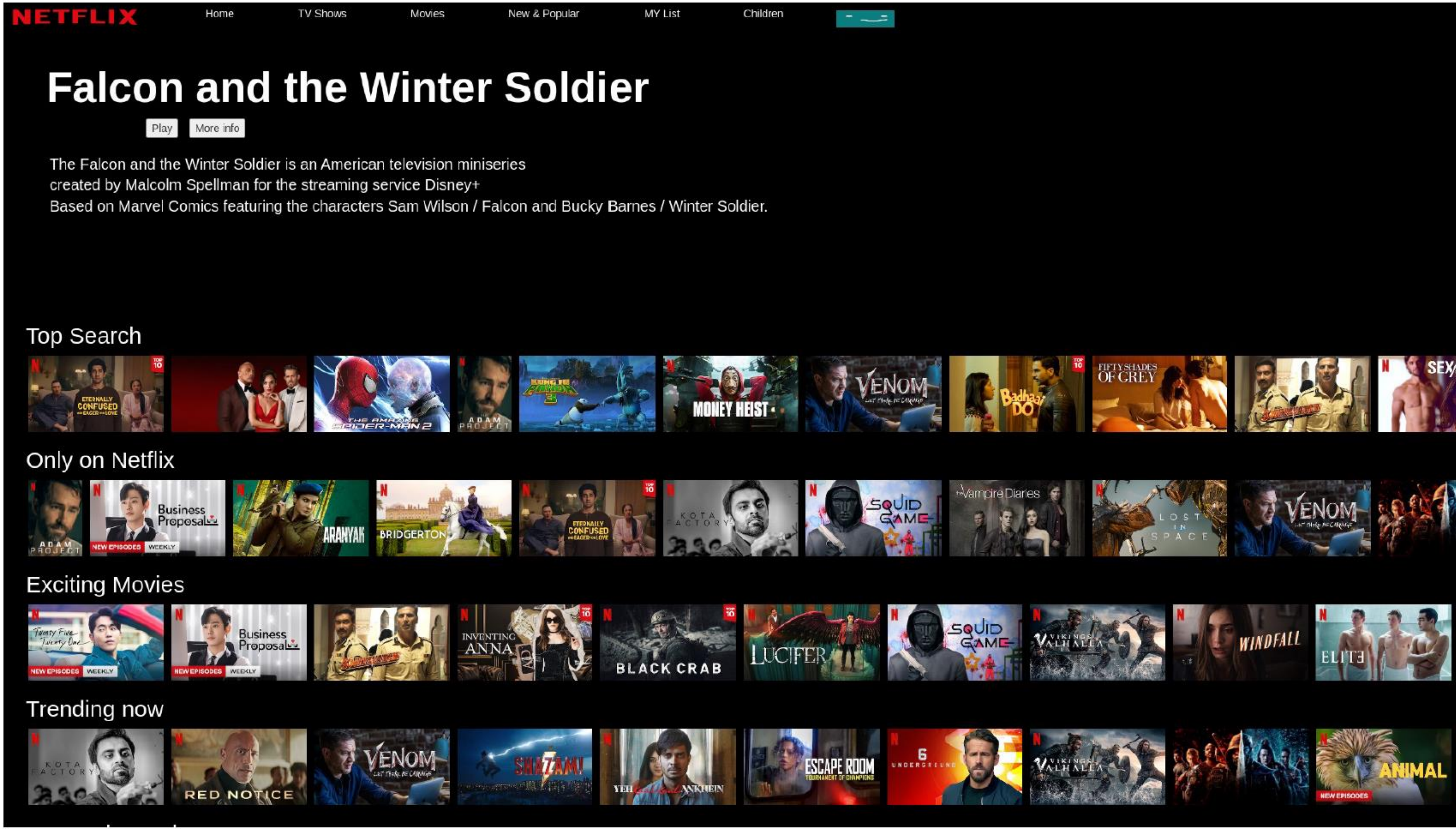
"On the screenshot you see the login page of a social media platform where users can share photos, videos and messages with their followers."

(larger image: here)

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

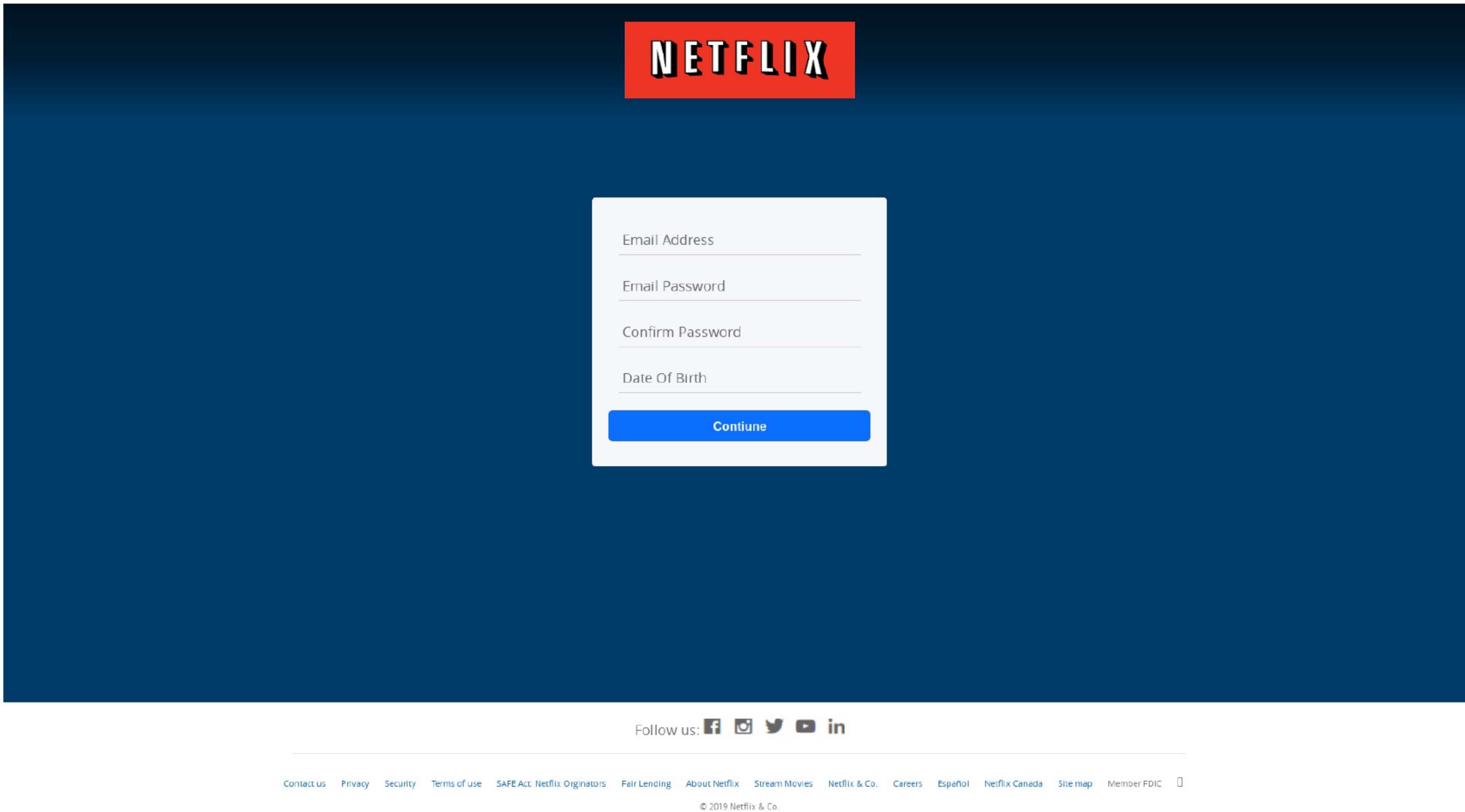
Fig. 3: Exemplary question (i.e., the first) in part II of our questionnaire. The screenshot refers to an adversarial webpage.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li



(a) Screenshot 10 (“moderate difficulty” to identify as phishing—by humans).

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

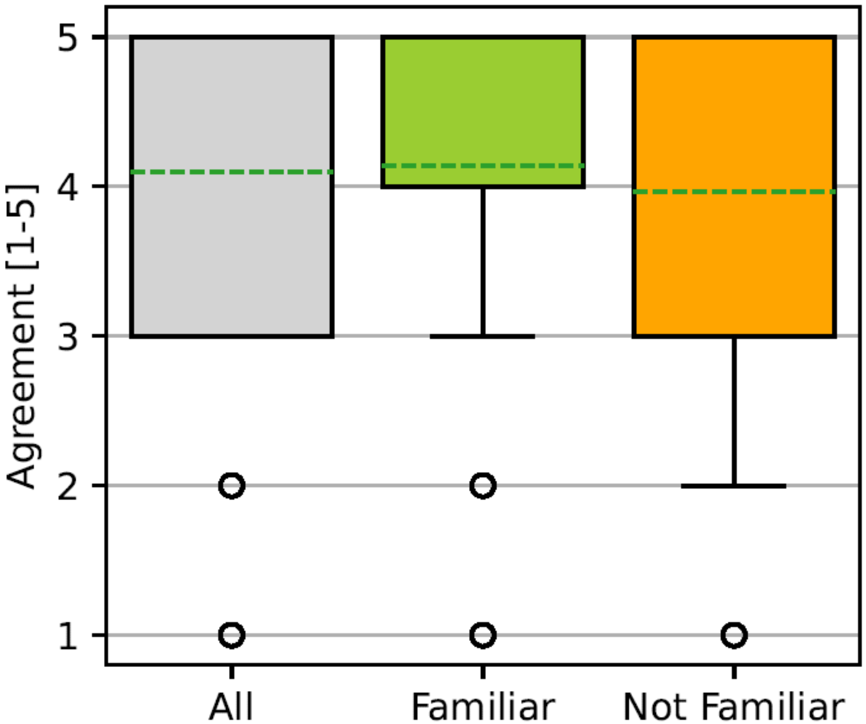


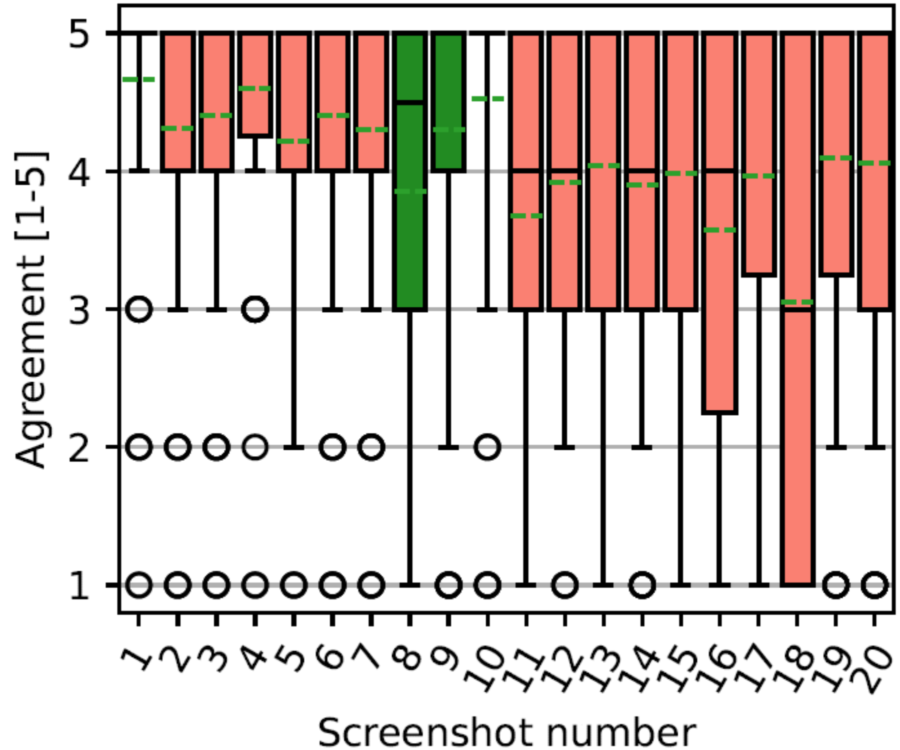
(b) Screenshot 18 ("very easy difficulty" to identify as phishing—by humans).

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What did we find? (1)

Higher agreement = higher likelihood of being deceived





TAKEAWAY. Most of our sample cannot recognize AW, and familiarity with a brand hinders the detection skills of users.

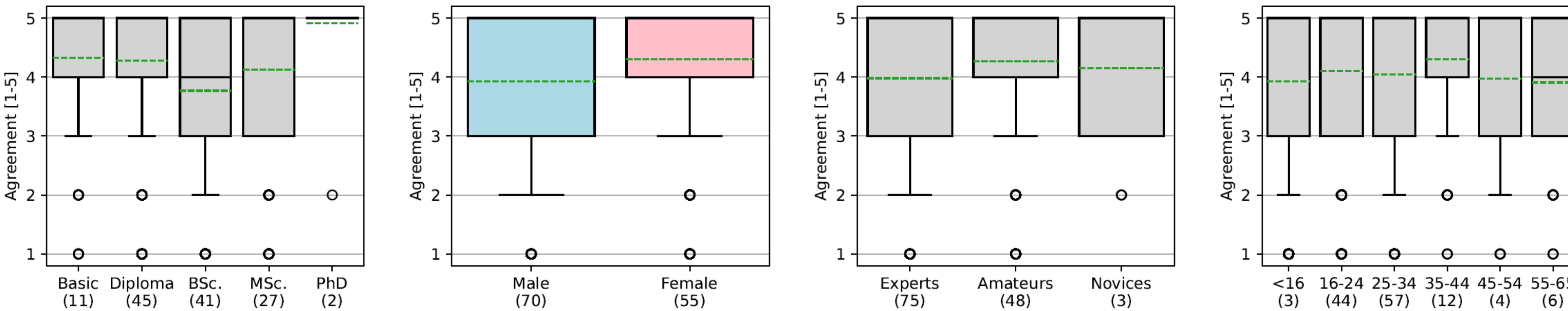
LIECHTENSTEIN

These claims are statistically significant ($p<0.05$)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What did we find? (2)

Higher agreement = higher likelihood of being deceived



(a) Education. (b) Gender. (c) Expertise with IT. (d) Age.

Fig. 5: Subgroup results. The figures report the aggregated ratings (for the 18 AW) of each subgroup (the x-axis shows the size of each subgroup).

- University graduates are more suspicious
- Female appear to be less suspicious than males
- IT experts are more skeptical than amateurs
- Age is not correlated with suspiciousness

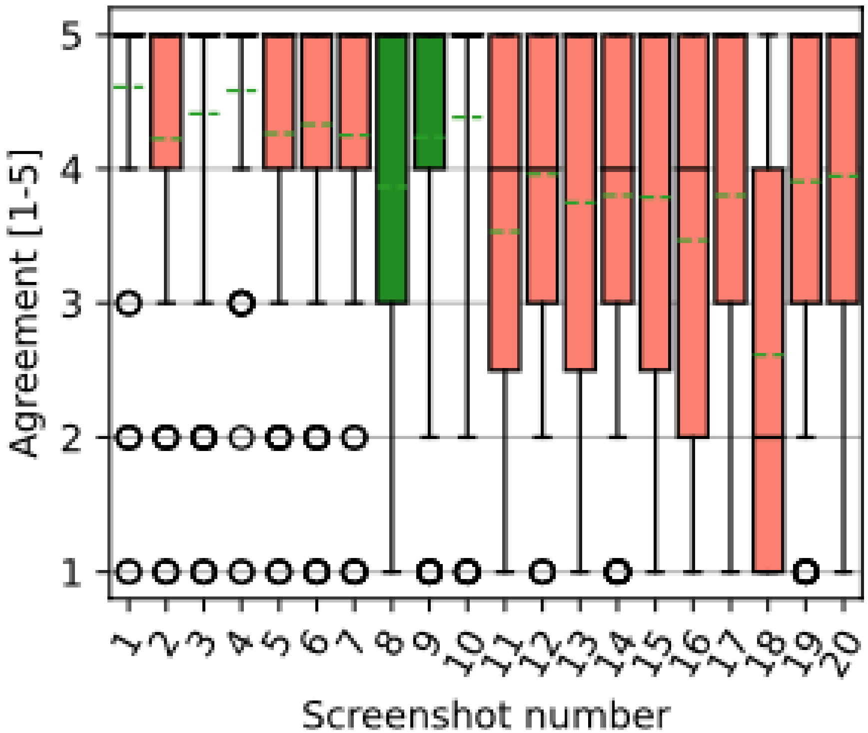
UNIVERSITÄT LIECHTENSTEIN

These claims are statistically significant ($p<0.05$)

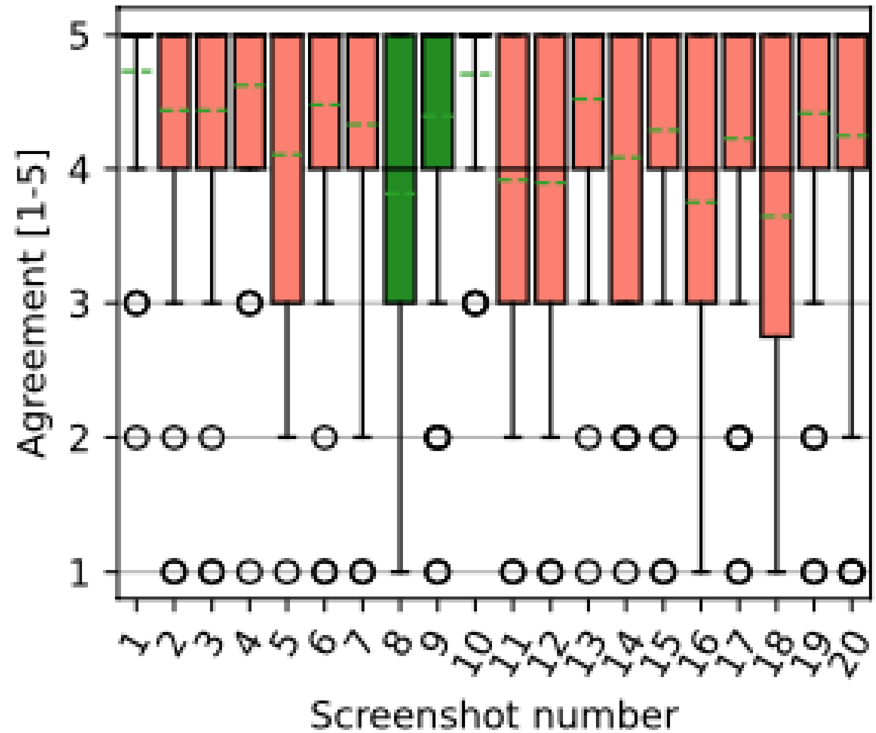
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What did we find? (3)

IT expertise influences the skepticism of participants



IT experts



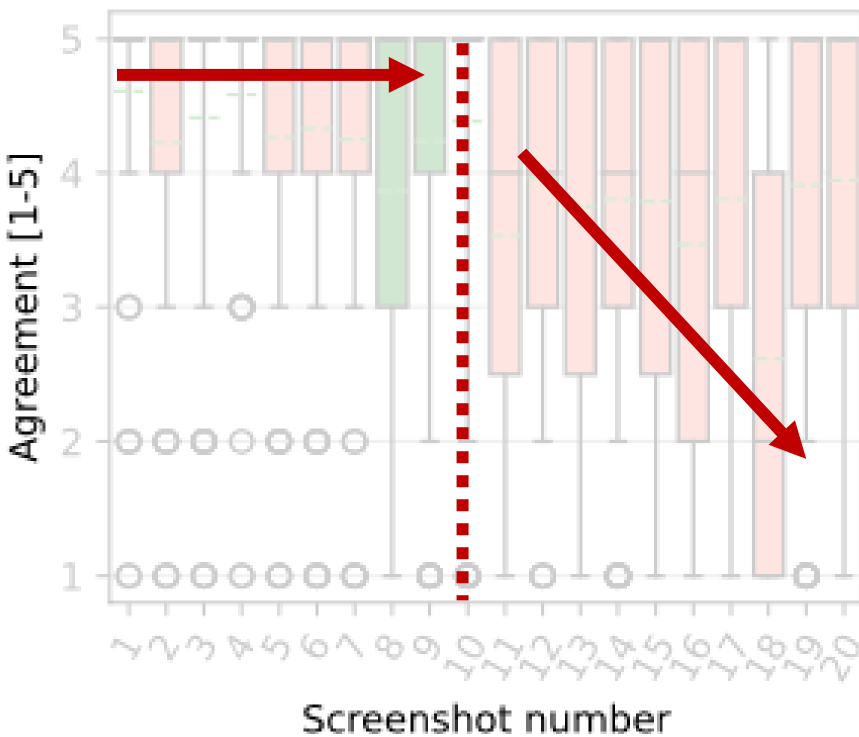
IT amateurs

UNIVERSITÄT LIECHTENSTEIN

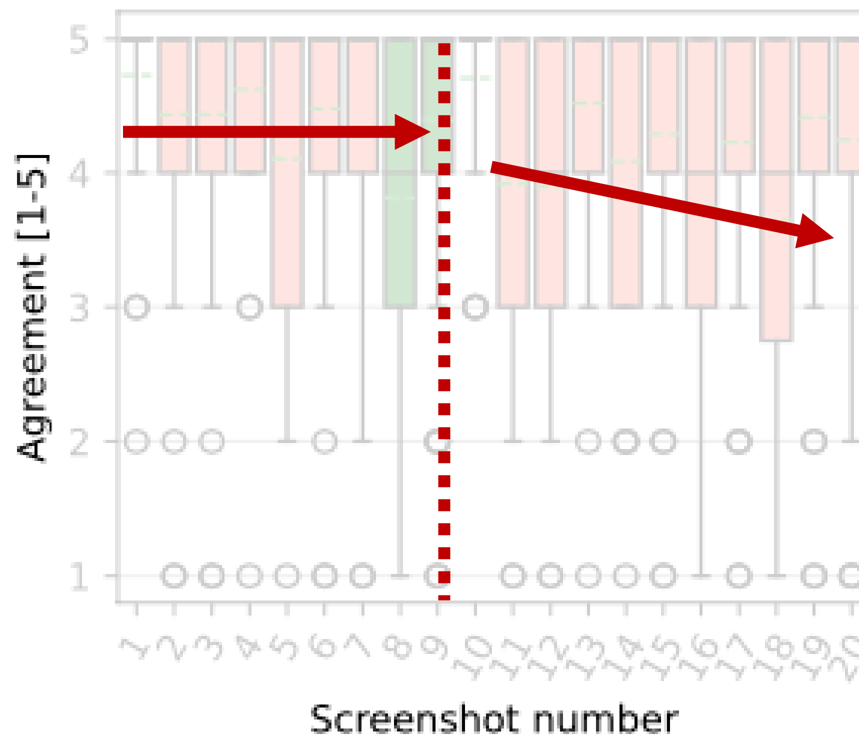
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What did we find? (3)

IT expertise influences the skepticism of participants



IT experts



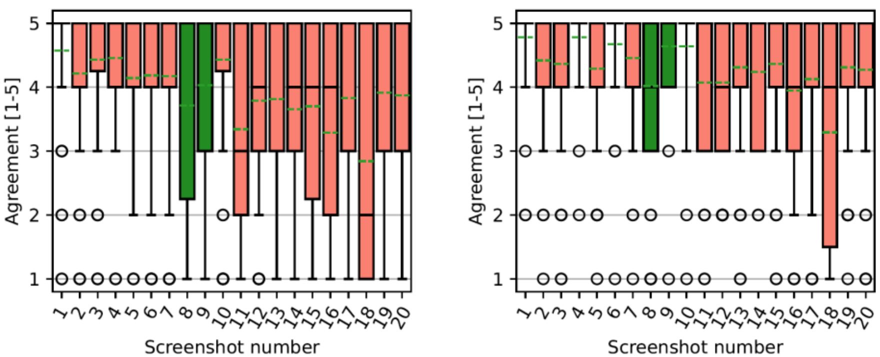
IT amateurs

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

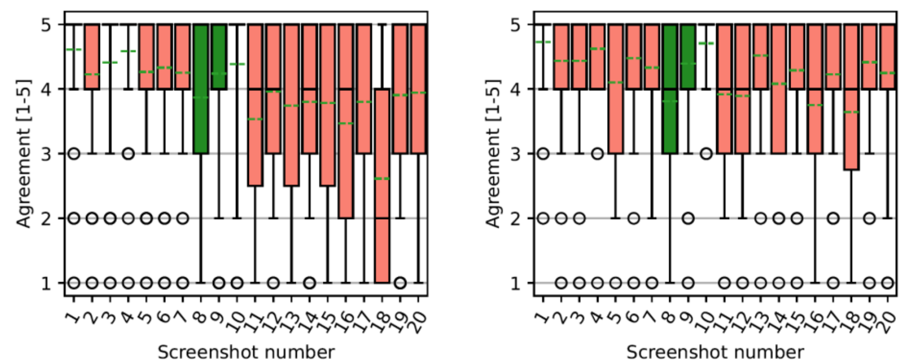
What did we find? (4)

Higher agreement = higher likelihood of being deceived



(a) Male (N=70). (b) Female (N=55).

Fig. 6: Individual screenshot ratings based on Gender.



(a) IT experts (N=75). (b) IT amateurs (N=48).

Fig. 7: Individual screenshot ratings based on Expertise with IT.

> **TAKEAWAY.** As participants advance in our questionnaire, they appear to become more suspicious.

UNIVERSITÄT LIECHTENSTEIN

These claims are statistically significant ($p<0.05$)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What do users think? (1)

- At the end of the questionnaire, we also asked each participant to provide some "explanations" for the skepticism on some screenshots.
- We analysed these through inductive coding (we devised a codebook)

	Altered Visual Logo
Screenshot #10	"because of the logo. It's squeezed together" "logo/branding looks fake. The font on the categories doesn't fit." "Logo is not on top right and everything is very distorted/compressed" "Looks fake. (Logo, layout)" "slightly different logo"
Screenshot #18	"wrong Netflix logo - fake" "wrong logo, it hasn't existed like this for years" "wrong logo" "I find the logo weird, but it seems to be the page for registration, so not login but registration if the logo is not fake" "different logo and different colors" "completely different logo"

Work in Progress

UNIVERSIT
LIECHTENS

TAKEAWAY. Several participants noticed some "common phishing elements" that can be acted upon (by practitioners) to improve existing PDS against (real) evasive webpages.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What do users think? (2)

- At the end of the questionnaire, we also asked each participant to provide some "explanations" for the skepticism on some screenshots.
- We analysed these through inductive coding (we devised a codebook)

Unusual Login Functionality and Style
N/A
"Screenshot looks more like password renewa" "completely different interface, Netflix doesn't use blue as much, generally different login and design" "the Netflix login page looks different in my opinion" "you can see the registration page not the login page" "the login page looks different than what I'm used to. I find a little confusing/different" "not login, but password change" "the registration page of Netflix that I know looks different"

Work in Progress

UNIVERSIT
LIECHTENS

TAKEAWAY. Several participants noticed some "common phishing elements" that can be acted upon (by practitioners) to improve existing PDS against (real) evasive webpages.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What do users think? (3)

- At the end of the questionnaire, we also asked each participant to provide some "explanations" for the skepticism on some screenshots.
- We analysed these through inductive coding (we devised a codebook)

Different style of text and font
"Looks a little distorted in the picture, not sure. May well be fake" "weird rendering and font" "Logo, Layout" "The interface of Netflix looks different. The "tabs" are arranged on the left, etc." "Wasn't exactly sure-the headings look different somehow (font & size)."
"modern login page looks different" "looks cheap, something is wrong there" "Layout is too old fashioned, today Netflix login looks different" "looks like a fake site" "outdated design" "too minimalistic if you don't know the site"

Work in Progress

UNIVERSIT
LIECHTENS

TAKEAWAY. Several participants noticed some "common phishing elements" that can be acted upon (by practitioners) to improve existing PDS against (real) evasive webpages.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Takeaways

Adversarial webpages are a problem in reality.

1. Investigating the human perception **is feasible**
2. To simulate a realistic setting, **avoid priming…**
3. …and **make it short!** (even when not primed, users become skeptical over time!

Complete alignment, however, is hard!

(and practitioners should lend a hand…)

Scientific Research ⟷ **Operational Practice** ⟷ **Human Factor**

UNIVERSITÄT LIECHTENSTEIN

Barcelona – November 16th, 2023

APWG Symposium on Electronic Crime Research

"Do Users fall for Real Adversarial Phishing?" Investigating the Human response to evasive Webpages

Ajka Draganovic, Savino Dambra, Xavier Aldana Iouit,
Kevin Roundy, Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

Avast

NortonLifeLock