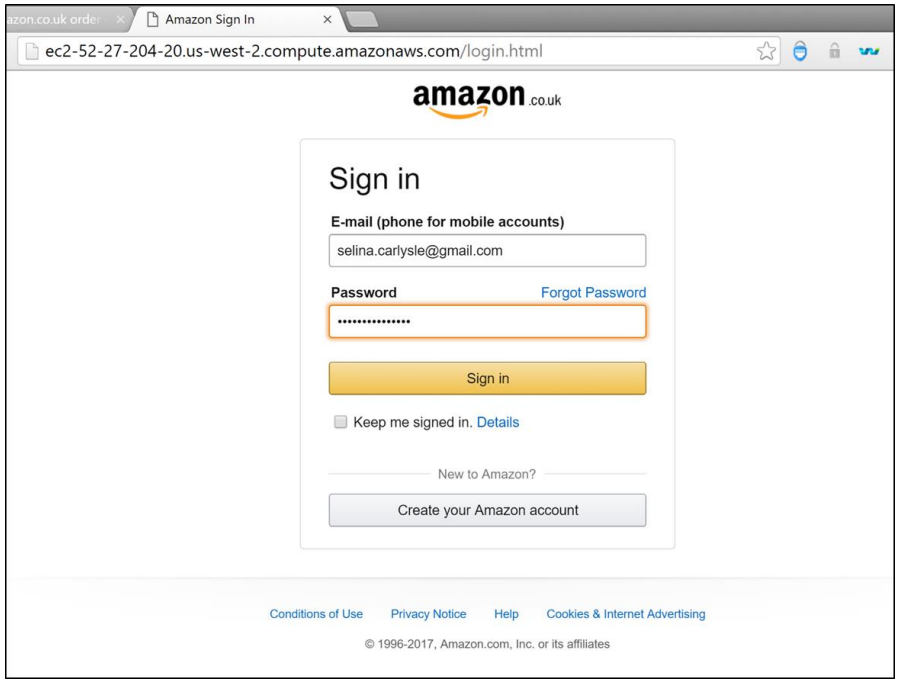
4 July 2023 - Security for all in an AI enabled society

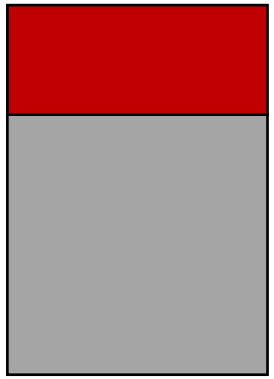
Cyber Hygiene in AI-enabled domestic life: A smart heating case study experiment

Professor George Loukas

University of Greenwich

From Human-as-a-Problem to Human-as-a-solution in cyber security – “old-school (social media, web, email cyber threats)





2 in 3 get it right wrong

Commercial cyber security products Vs. humans

	AntiVirus						
Attack	A1	A2	A3	A4	A5	A6	A7
1.1	✗	✗	✗	✗	✗	✗	✗
1.2	✓	✗	✗	✗	✗	✗	✗
2.1	✗	✗	✗	✗	✗	✗	✗
2.2	✗	✗	✗	✗	✗	✗	✗
3.1	✗	✗	✗	✗	✗	✗	✗
3.2	✗	✗	✗	✗	✗	✗	✗

	Browsers						
Attack	B1	B2	B3	B4	B5	B6	B7
1.1	✗	✗	✗	✗	✗	✗	✗
1.2	✓	✗	✗	✗	✗	✗	✗
2.1	✗	✗	✗	✗	✗	✗	✗
2.2	✗	✗	✗	✗	✗	✗	✗
3.1	✗	✗	✗	✗	✗	✗	✗
3.2	✗	✗	✗	✗	✗	✗	✗

	Platform		
Attack	P1	P2	P3
1.1	-	-	✗
1.2	-	✗	✓
2.1	✗	-	✗
2.2	✗	-	✗
3.1	-	-	✗
3.2	-	-	✗

	AntiVirus						
Attack	A1	A2	A3	A4	A5	A6	A7
2.1A	✗	✗	✗	✗	✗	✗	✗
4.1	✗	✗	✗	✗	✗	✗	✗
4.2	✗	✗	✗	✗	✗	✗	✗
5.1	✗	✗	✗	✗	✗	✗	✗
5.2	✓	✗	✗	✗	✗	✗	✗

	Browsers						
Attack	B1	B2	B3	B4	B5	B6	B7
2.1A	✗	✗	✗	✗	✗	✗	✗
4.1	✓	✗	✗	✗	✗	✗	✗
4.2	✗	✗	✗	✗	✗	✗	✗
5.1	-	-	-	-	-	-	-
5.2	-	-	-	-	-	-	-

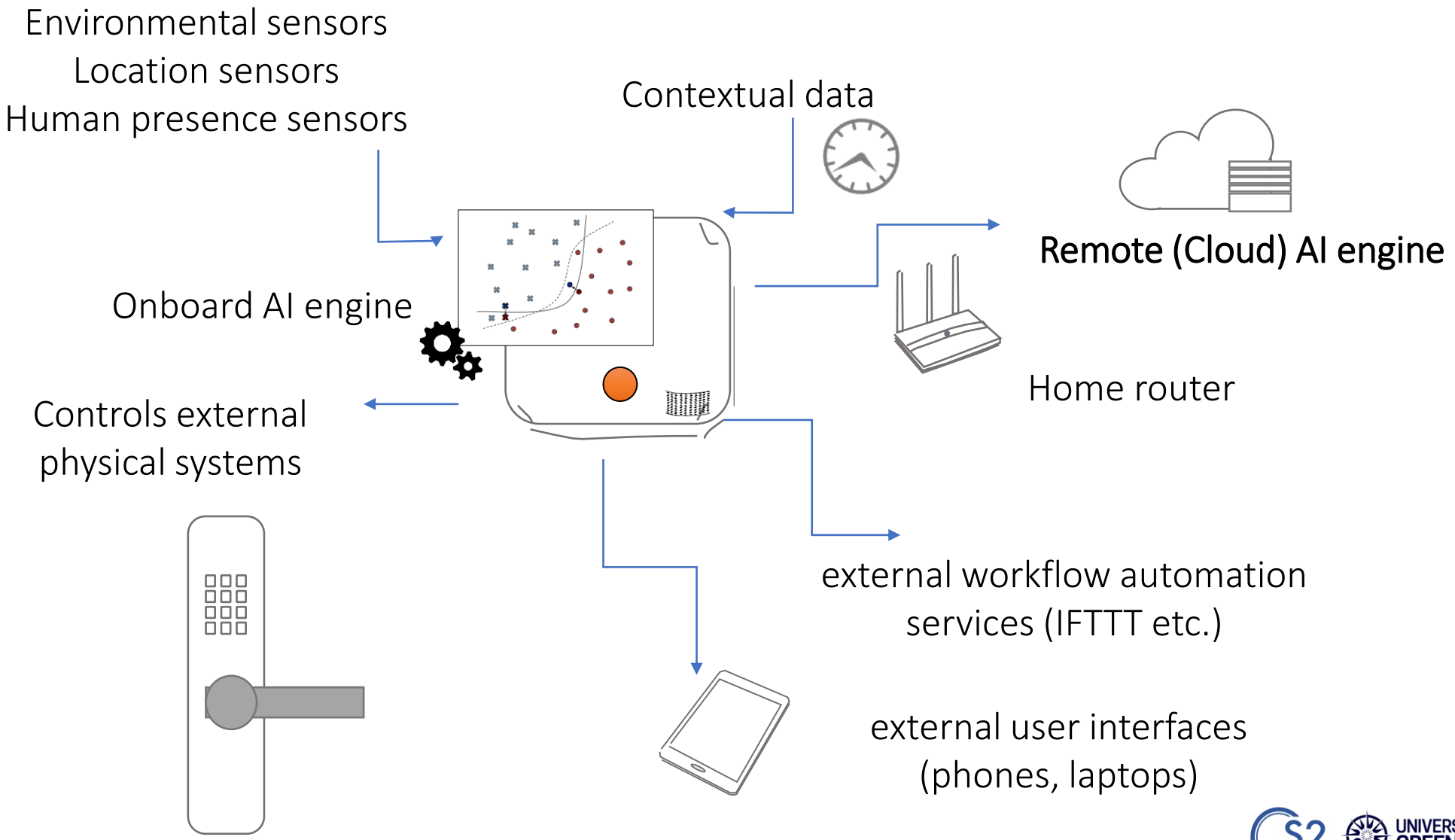
	Platform	
Attack	P1	P3
2.1A	✗	-
4.1	✗	-
4.2	✗	-
5.1	-	✗
5.2	-	✗

A	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	H21	H22	H23	H24	H25	H26
1.1	.49	.72	.54	.55	-	.59	✗	.50	-	-	.70	.59	.48	-	.53	-	-	.33	.46	-	-	.46	.54	.70	.79	.66
1.2	-	.66	.53	-	-	-	✓	.26	-	-	-	-	.34	-	.18	-	-	.44	-	-	-	-	.17	-	.71	-
2.1A	-	-	-	-	-	.60	✓	-	-	-	.58	-	-	-	.48	-	-	-	-	.49	.30	-	-	-	-	-
2.1B	.51	-	-	-	-	-	✓	.51	-	-	.58	.43	-	-	.48	-	-	.32	-	.49	-	-	-	-	-	-
2.2	.51	-	-	-	-	-	-	-	-	-	-	-	-	-	.48	-	-	.32	-	-	-	-	-	-	-	-
3.1	-	-	.54	.55	-	.60	✗	.63	-	.51	.70	.59	-	.56	.56	.46	.49	.33	.46	-	-	-	.54	.70	-	.66
3.2	-	-	-	-	-	-	.62	-	-	.49	-	-	-	.48	-	-	.26	.44	.46	-	-	-	-	.49	-	.67
4.1	-	-	-	.63	-	.56	-	-	-	.53	.72	.43	-	-	-	-	-	-	-	.49	-	-	-	-	-	-
4.2	-	-	-	-	-	-	-	-	-	.56	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5.1	-	.67	-	.61	✗	.63	✗	.39	.35	-	.65	.52	-	-	.49	-	-	.45	.47	-	.30	.37	.66	.44	-	.68
5.2	-	-	-	-	✗	-	✓	-	.35	-	-	-	-	-	-	-	-	-	.47	-	.57	-	.68	.72	-	.64

Heartfield, R. and Loukas, G. (2018) Detecting semantic social engineering attacks with the weakest link. Computers & Security, Elsevier.

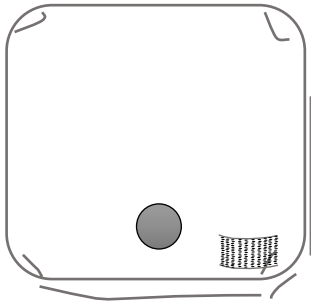
S2 Centre for Sustainable Cyber Security | UNIVERSITY OF GREENWICH

Is the concept still applicable in AI-enabled society?



Is the concept still applicable in AI-enabled society?

“It’s playing up again”



Is the concept still applicable in AI-enabled society?

There are many reasons to want it to be.

- People **feeling more in control** of their own devices and systems
- **Extra layer of defence** especially for new and unknown threats to AI
- People **will not trust AI blindly**

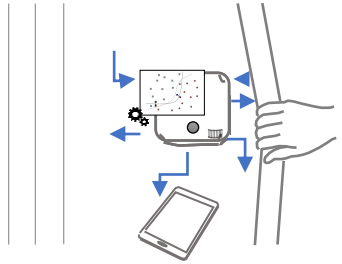
Is the concept still applicable in AI-enabled society?

In CHAI, we explore this from four angles

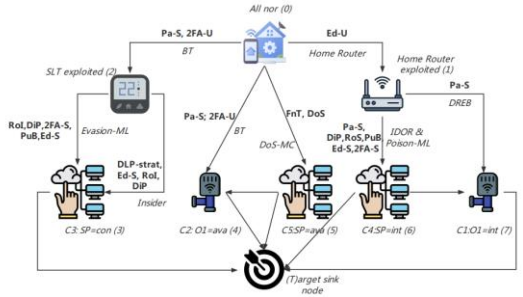
Cyber hygiene for AI users (+ Training)

Perception (What to notice)
Detection (How to diagnose)
Response (How to act)

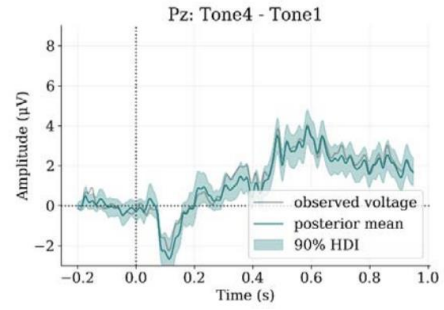
AI transparency for non-experts



Optimisation of AI measures for the service provider



Experimental study of the neuroscience dimensions



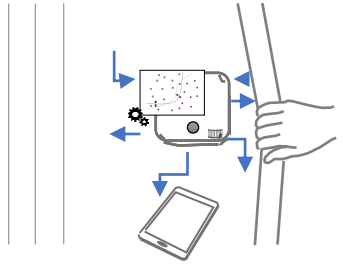
Is the concept still applicable in AI-enabled society?

The first two are linked directly

Cyber hygiene for AI users (+ Training)

Perception (What to notice)
Detection (How to diagnose)
Response (How to act)

AI transparency for non-experts



Insights from literature

Cyber hygiene

Cyber hygiene clearly leads to online cyber safety, encompassing what we think, how we think, and what we do online

Vishwanath, A., Neo, L.S., Goh, P., Lee, S., Khader, M., Ong, G. and Chin, J., 2020. Cyber hygiene: The concept, its measure, and its initial tests. Decision Support Systems, 128.

In cybersecurity, humans are considered a problem to be controlled. this robs organisations of their human agents' ability to make a contribution to cyber security.

Zimmermann, V. and Renaud, K., 2019. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. International Journal of Human-Computer Studies, 131, pp.169-187.

AI transparency for non-experts

Gaining even limited knowledge of the inner workings of AI approaches is important for non-experts to develop trust.

Ayobi, A., Stawarz, K., Katz, D., Marshall, P., Yamagata, T., Santos-Rodríguez, R., Flach, P. and O'Kane, A.A., 2021. Machine learning explanations as boundary objects: how AI researchers explain and non-experts perceive machine learning.

AI risk communication suffers from disconnect between practitioners' and researchers' mental models of threats.

Bieringer, L., Grosse, K., Backes, M., Biggio, B. and Krombholz, K., 2022. Industrial practitioners' mental models of adversarial machine learning. In SOUPS 2022.

Explainable AI can help secure human-interactive robots

Roque, A. and Damodaran, S.K., 2022. Explainable AI for Security of Human-Interactive Robots. International Journal of Human–Computer Interaction, 38(18-20), pp.1789-1807.

S2 Centre for Sustainable Cyber Security

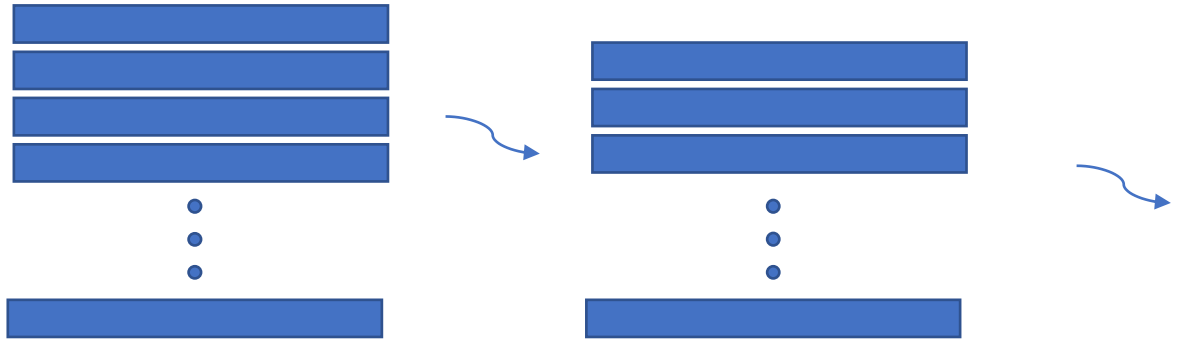
UNIVERSITY OF GREENWICH

Existing Cyber Hygiene and AI-enabled smart homes

Analysed applicability of 105 IoT security recommendations provided by NCSC, ENISA, etc. in terms of **applicability to prevention**, **perception**, **detection and response for AI-IoT**, **population** (expertise required), **intervention** (time required), and **scientific support**.

All recommendations
105

Applicable for common AI-IoT with no heavy expertise or time
75



Substantial scientific evidence
only 5
(**all preventive** and **only indirectly useful**)

PR-1: Turn off service when not needed
PR-2: Change all default passwords or PINs
PR-3: Use different passwords for different systems
PR-4: Do not use passwords that are based on personal information that can be easily accessed or guessed
PR-5: Do not use passwords that can be found in any dictionary of any language

S2 Centre for Sustainable Cyber Security
UNIVERSITY OF GREENWICH

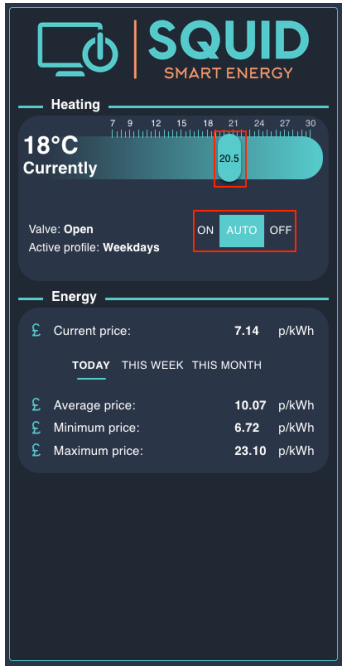
A smart heating case study (AI-controlled radiator valve)

Aiming for simplest possible scenario to explain to a non-expert:

Schedule learning for **radiator valve setpoint Vs. price** at different times of the day.

Bayesian linear regression

Separate model for each of five profiles (mornings, weekdays/weekends, evenings, nights)



https://github.com/chai-project

S2 Centre for Sustainable Cyber Security
UNIVERSITY OF GREENWICH

AI Transparency measures in Squid application (1/2)

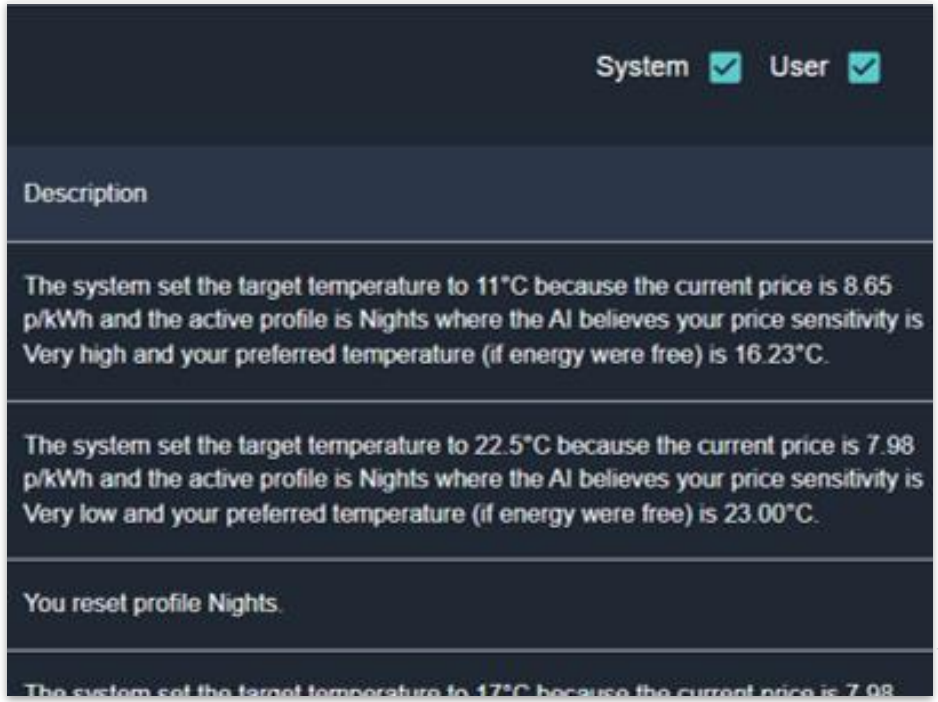
Aim: To provide more context on AI decisions than is usually available to the users, so as to help them notice AI misbehaviour

Preferred temp at price 0 and price sensitivity

Log of notifications

Example:

*"The system set the target temperature to **11 C** because the current price is **8.65 p/kWh** and the active profile is **Nights** where the AI believes your price sensitivity is **Very high** and your preferred temperature (if energy were free) is **16.23 C**."*



UNIVERSITY OF GREENWICH

AI Transparency measures in Squid application (2/2)

Taxonomy of attacks on AI-enabled smart homes

AV: Attack vector

CI: Cyber impact

IAI: Impact on AI operation

IDL: Impact on domestic life

AC-SC1: Supply Chain

AC-CS: Control Software

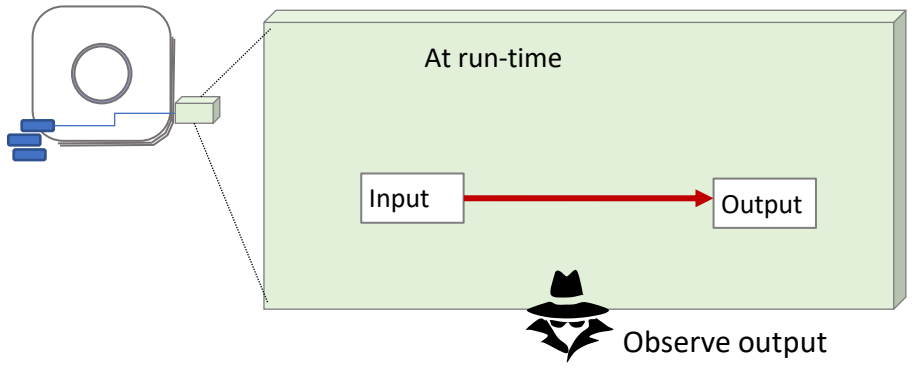
AC-CM: Communication Medium

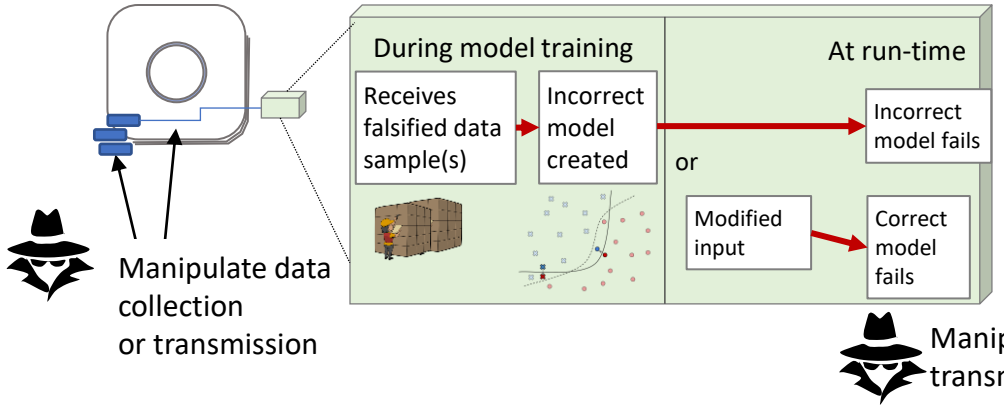
AC-SC2: Sensory Channel

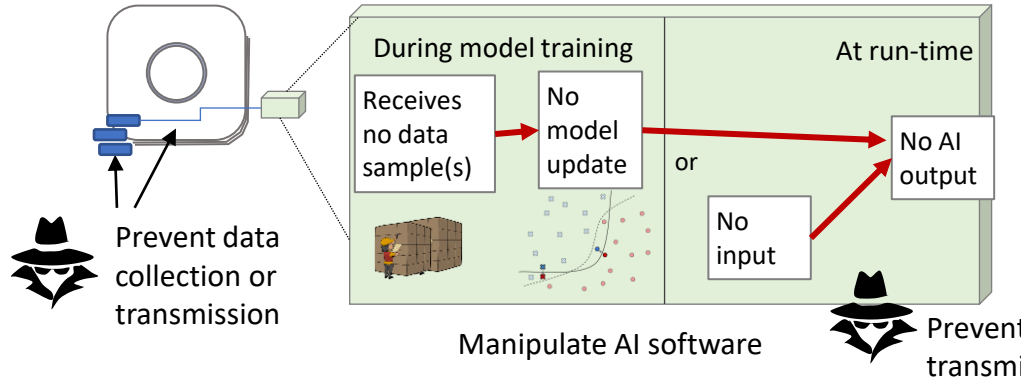
CI-C: Confidentiality

CI-I: Integrity

CI-AL Availability







IAI – II: Infer AI input

IAI-MD: Manipulate AI decision

IAI-PD: Prevent AI decision

IDL-IT: Information theft

IDL-APC: Assisting in physical crime

IDL-H: Harassment

IDL-PD: Physical damage

IDL-F: Financial

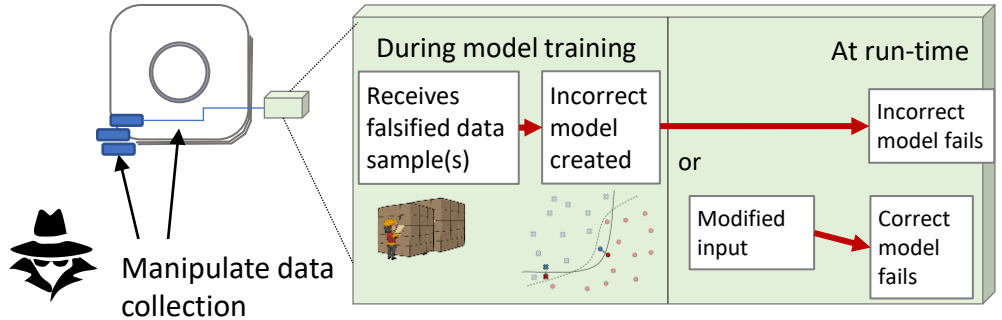
Attacks in CHAI experiment

AV: Attack vector | CI: Cyber impact | IAI: Impact on AI operation | IDL: Impact on domestic life

AC-CS: Control Software

AC-CM: Communication Medium

CI-I: Integrity



IAI-MD: Manipulate AI decision

IDL-H: Harassment

IDL-PD: Physical damage

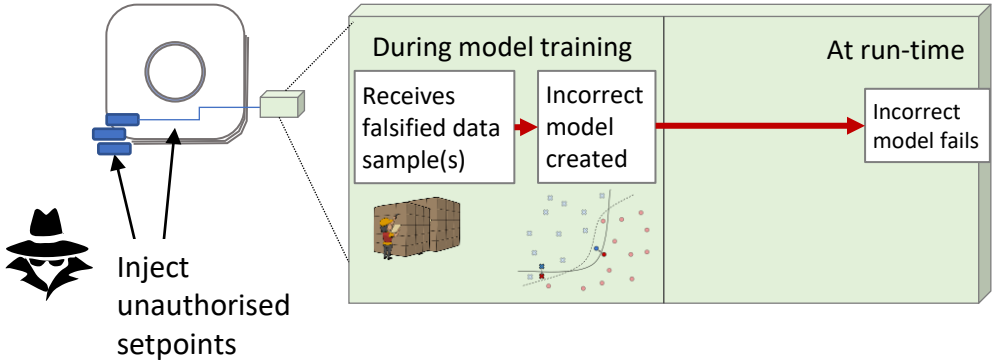
Focused on integrity attacks because they leave physical and digital traces observable by non-experts

Attacks in CHAI experiment (1/3)

AC-CS: Control Software

AC-CM: Communication Medium

CI-I: Integrity

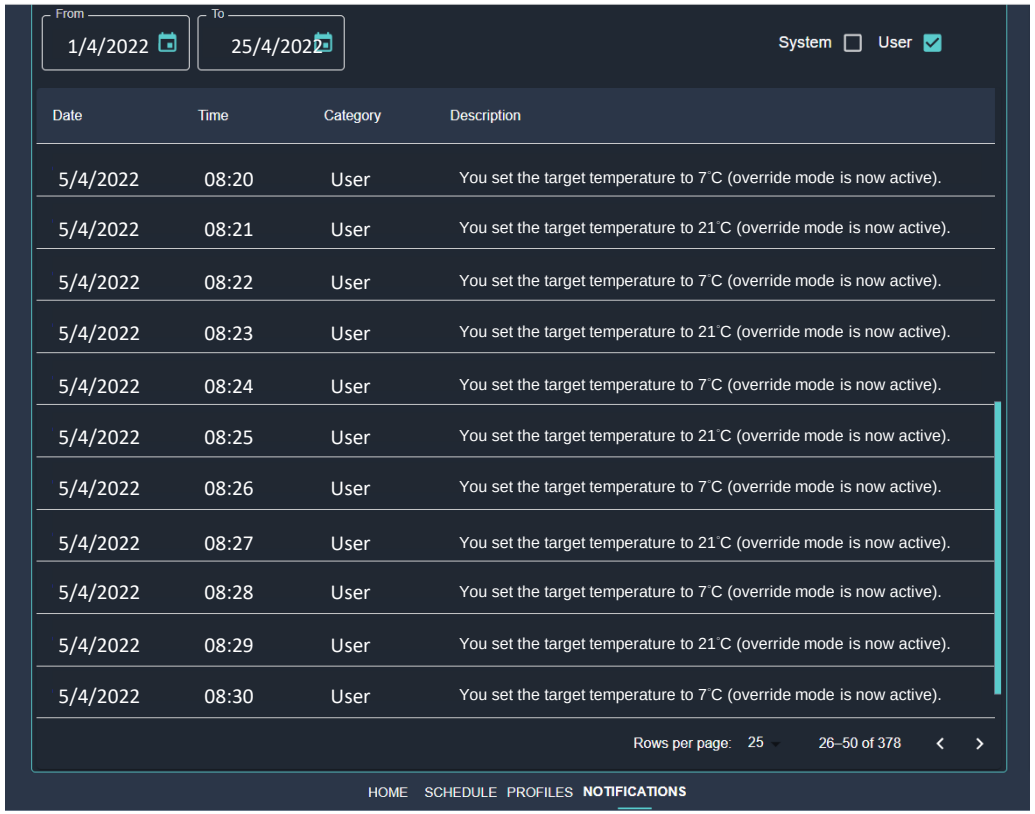


IAI-MD: Manipulate AI decision

IDL-H: Harassment

IDL-PD: Physical damage

Simple poisoning to modify price sensitivity or preferred temp at price 0.
Method: *Unauthorised setpoint injections*

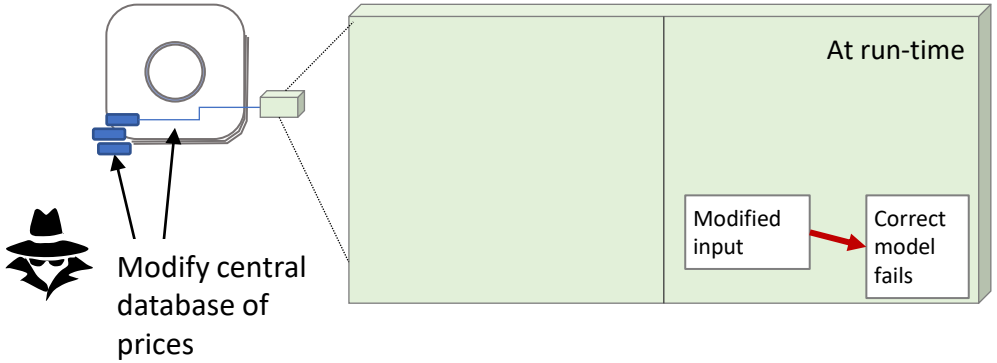


Attacks in CHAI experiment (2/3)

AC-CS: Control Software

AC-CM: Communication Medium

CI-I: Integrity



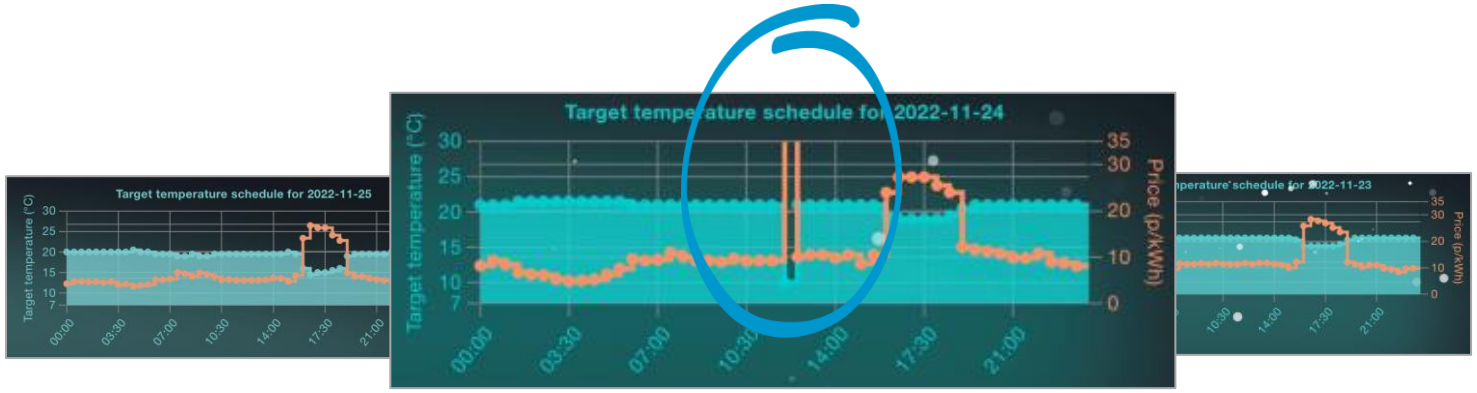
IAI-MD: Manipulate AI decision

IDL-H: Harassment

IDL-PD: Physical damage

Price manipulation to affect AI setpoints at scale

Method: Central database modified for a period of time



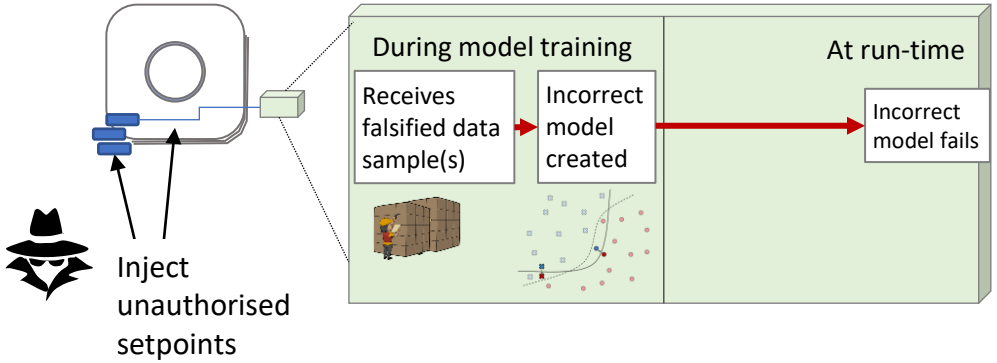
Day 1 **Day 2** **Day 3**

Attacks in CHAI experiment (1/3)

AC-CS: Control Software

AC-CM: Communication Medium

CI-I: Integrity

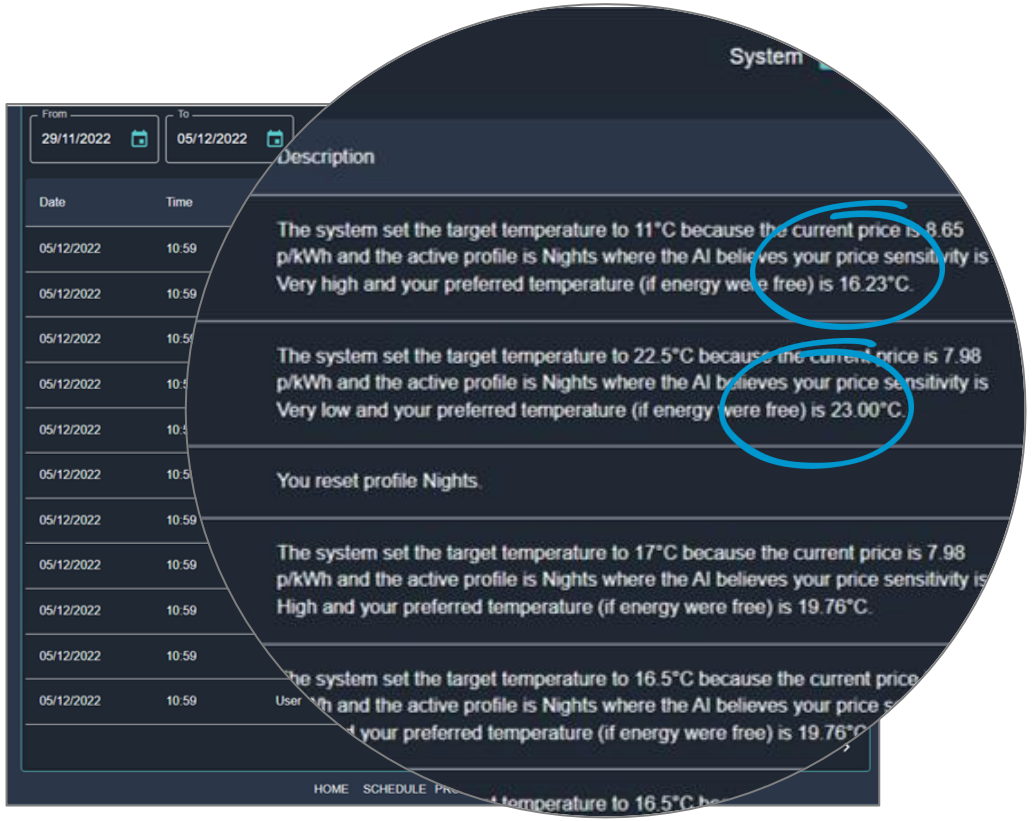


IAI-MD: Manipulate AI decision

IDL-H: Harassment

IDL-PD: Physical damage

Stealthy poisoning to modify price sensitivity or preferred temp at price 0.
Method: Unauthorised setpoint injections and deletion of log entries

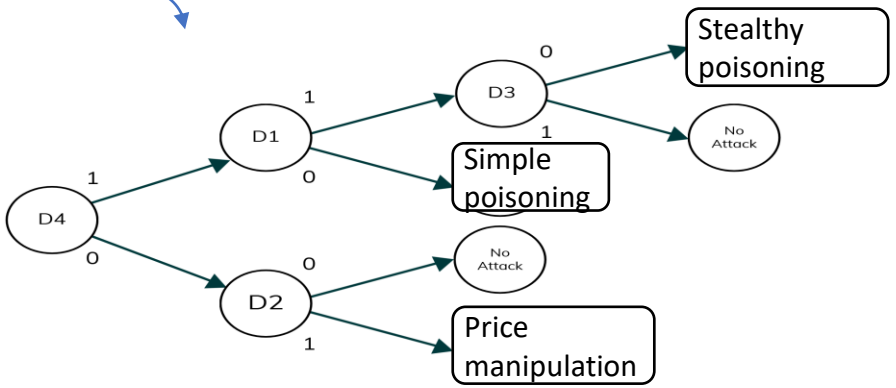


Cyber hygiene for perception, detection and response
A simple diagnostic graph method

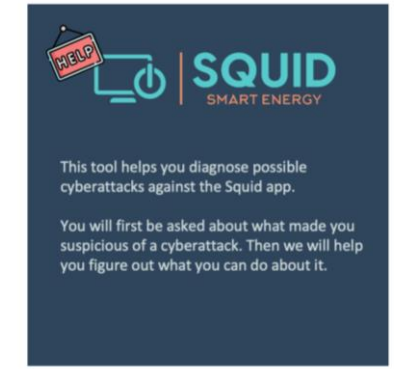
Truth table of observable indicators -> Binary decision diagram for non-experts

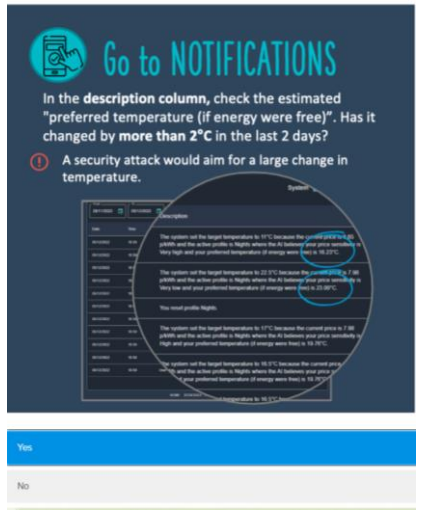
	Diagnostics (persistent facts)				Perception (temporary observations that trigger suspicion)		
	D1	D2	D3	D4	P1	P2	P3
Simple poisoning	0	0	0	1	1	1	1
Input manipulation	1	1	0	0	0	0	0
Stealthy poisoning	1	0	0	1	1	1	1
Normal rapid setpoint changes	1	0	1	1	1	1	1

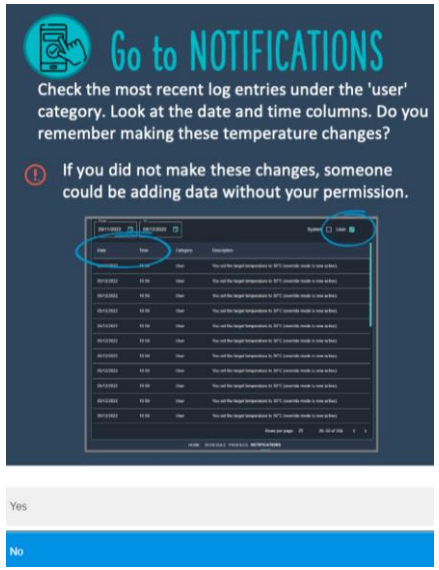
D1: all user entries are recognised
D2: dramatic price change
D3: several recent user entries
D4: significant preferred temperature at price 0 change
P1: frequent buzz
P2: room too warm/cold
P3: unusual price sensitivity



Cyber hygiene for perception, detection and response
Diagnostic graph UI and indicative content







You may have detected an AI Poisoning Attack

It seems that a number of temperature changes have been entered into the system over a short period of time

These entries have become part of your current profile. Squid is heating the radiator according to this new data

WHAT TO DO NEXT

- ✓ **Look at the problematic log description and identify which profile has been breached. This needs to be reset**
- ✓ Go to *Profiles*. Select the name of the **breached profile** from the top left menu and **click on the *Reset this Profile*** button on the top right-hand side

S2 Centre for Sustainable Cyber Security
UNIVERSITY OF GREENWICH

Cyber hygiene for perception, detection and response
Experimental setup

19 participants from 10 households

<u>Schedule:</u>

Training on Squid

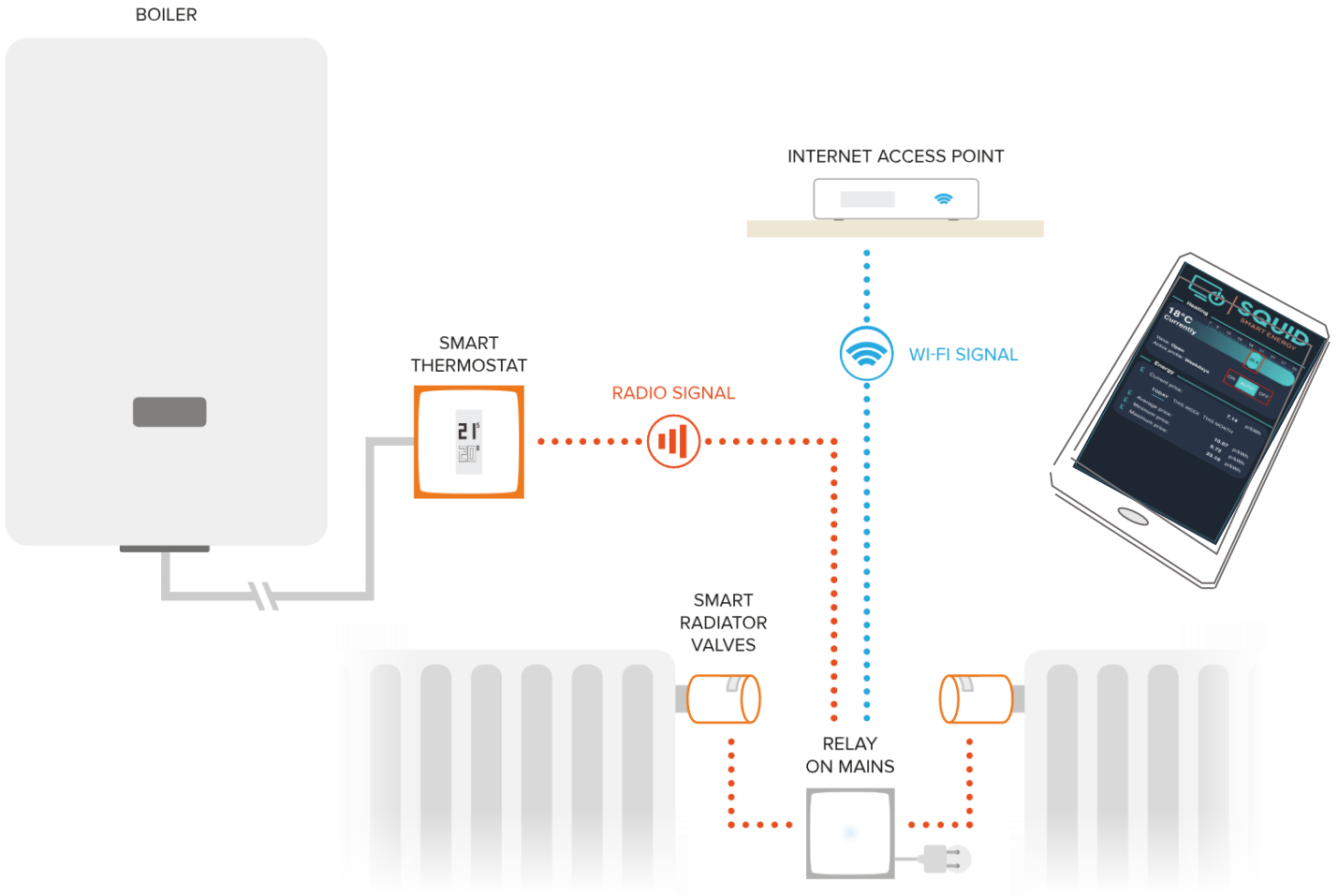
1 month of familiarisation

Training on one attack

1 month of attacks (2 x 3 attacks each)

<u>Data collection</u>:

Reporting diary for each incident suspected
Weekly interviews with each household
Squid usage data
Diagnostic tool usage data



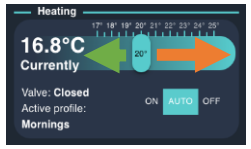
Cyber hygiene for perception, detection and response Training

1 hour session including slides and use of the diagnostic tool on the Simple poisoning attack.
No training on the AI transparency features or other attacks

Perception

What signs did you identify?

APP BEHAVIOUR

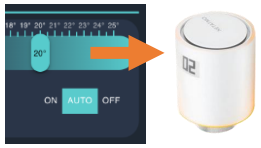


The target temperature on the app has **changed** a lot in the past 30 mins

ENVIRONMENT

The room feels a lot **warmer** than I would expect

DEVICE FUNCTION

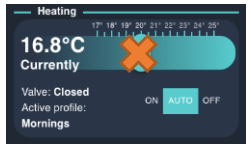


After increasing the target temp it took a **long time** for the valve to turn on

SOUNDS

I heard the radiator valve **clicking** on/off more often than normal

APP BEHAVIOUR



The target temperature on the app has **not changed** recently

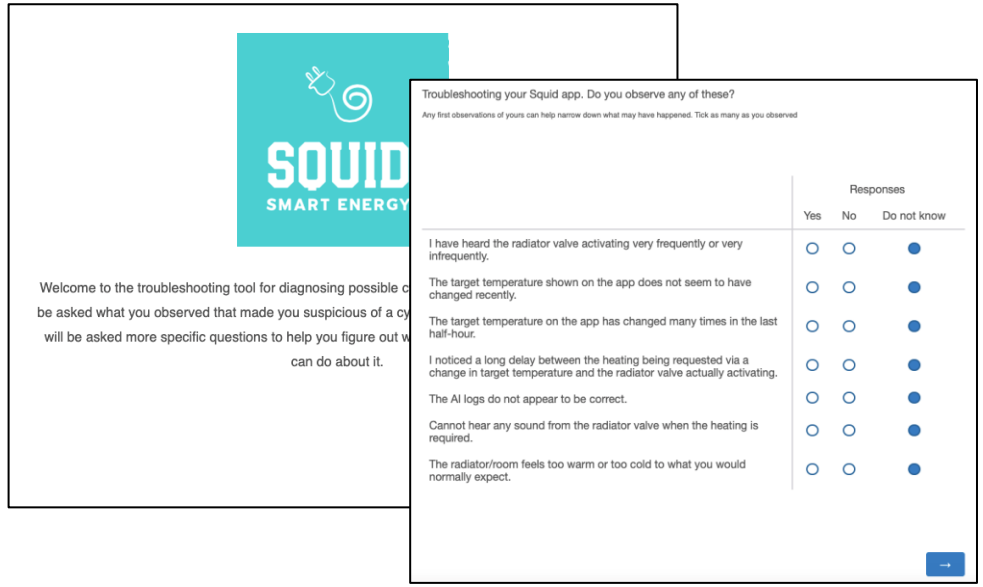
APP BEHAVIOUR

The **AI logs** do not appear to be correct

SOUNDS

There is **no sound** from the radiator valve when the heating turns on

Detection and response



S2 Centre for Sustainable Cyber Security
UNIVERSITY OF GREENWICH

Cyber hygiene for perception, detection and response
Experimental results 1 – Human as an AI threat sensor

Groundtruth

Perception

	Positive	Negative
Positive	47 (78%)	8 (4%)
Negative	13 (22 %)	212* (96%)

f1 score = 0.86

Groundtruth

Detection

	Positive	Negative
Positive	31 (60%)	2 (25%)
Negative	13 (40%)	6 (75%)

f1 score = 0.73

Response

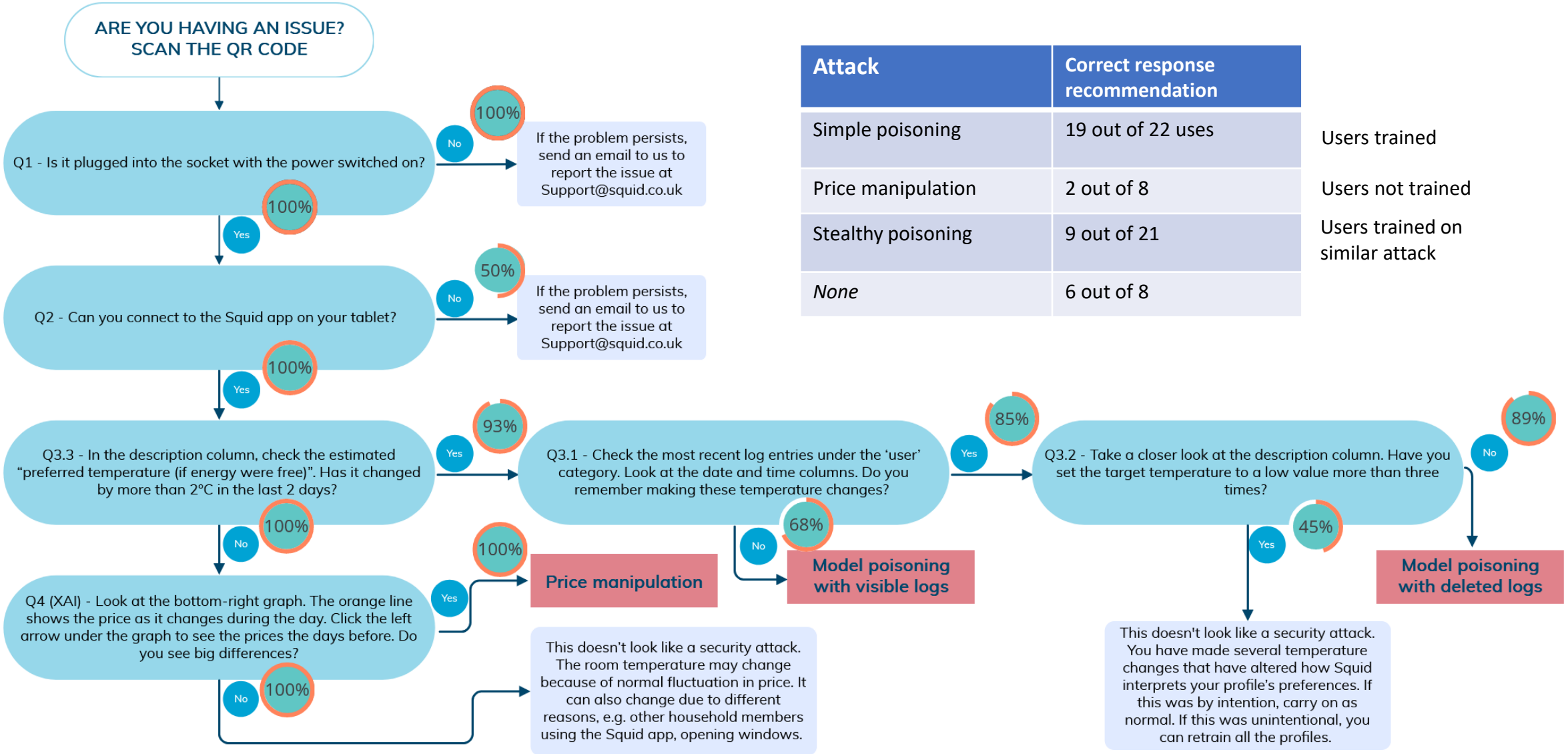
Groundtruth

	Attack	Non-attack
Correct Action	39 (78%)	6 (86%)
Incorrect Action	11 (22%)	1 (14%)

Also, no obvious correlation between level of engagement (number of interactions) and performance as human sensors

* True negative is a day where non-attack occurred and no attack was perceived

Cyber hygiene for perception, detection and response
Experimental results 3 – diagnostic tool



Attack	Correct response recommendation
Simple poisoning	19 out of 22 uses
Price manipulation	2 out of 8
Stealthy poisoning	9 out of 21
None	6 out of 8

Users trained

Users not trained

Users trained on similar attack

Initial insights

Existing cyber hygiene for AI-IoT is only preventive and not AI-specific.

We developed a domain-specific example set of perception, detection and response guidance, supported by partial training.

Perception accuracy was high (f1 score 0.86)

Detection worked well only for the simple attack we had already trained the participants. Much less so for the rest.

No correlation between number of interactions and human sensor performance

and next step

Test these initial insights quantitatively in lab-based experiment

Compare different AI explainability approaches

Assess practicality on different AI application areas

S2 Centre for Sustainable Cyber Security

UNIVERSITY OF GREENWICH

Thank you

UNIVERSITY OF GREENWICH