Doing cybersecurity at home: a human-centred approach for mitigating attacks in AI-enabled home devices

Abstract

AI-enabled devices are increasingly introduced in the home context and cyber-attacks targeting their AI component are becoming more frequent. Moving away from seeing the user as the problem to recognising the user as part of the solution, our research reports on a novel cybersecurity intervention (comprising Explainable AI features, assisted remediation) designed to support users to identify, diagnose and mitigate cyber-attacks on the AI component of their smart devices. We carried out a case study of a bespoke smart heating device inclusive of this intervention and conducted fieldwork with ten households who experienced simulated integrity cyber-attacks over a month. Our research contributes an understanding of how to design AI-enabled devices and their ecosystems to support users to perceive integrity cyber-attacks, offering new considerations for intervention design that exploits multimodal indicators and supports users to troubleshoot themselves the causes as well as actions of cyber-attacks. Contributing to the growing area of human-centred cybersecurity, we evidence the distinctive challenges users face when evaluating integrity attacks on the AI component in the home context.

Keywords

AI-enabled devices; smart devices; smart heating; human-centred; indicators; home; assisted remediation; cybersecurity; cyber-attacks; field study

1 Introduction

Connected smart devices are becoming a common and important feature of domestic life (Huijts et al., 2023), with promises of increased comfort and efficiency in the management of household tasks (Mennicken et al., 2014). Examples of the convenience offered by smart home devices can include ‘smart thermostats’ that ensure the home is at a desirable temperature when occupants arrive from work, ‘smart voice assistants’ that play one’s favourite music, or ‘smart locks’ that recognise the face of the dog walker to let them in. The everyday functionalities of these and other devices are often powered by artificial intelligence (AI), a trend set to continue as evidenced by the recent boom in AI. However, although widespread, the incorporation of AI in home technology devices confronts users with emerging cybersecurity challenges that remain underexplored. In contrast to their design intentions to make everyday life more efficient, comfortable, or enjoyable (Jensen et al., 2018b), these *AI-enabled devices* are also vulnerable to cyber-attacks owing to a combination of their integrated connectivity, multiple attack surfaces, opaque device behaviours, and the lack of standardised security legislation (Kuzlu et al., 2021).

In this paper, we are concerned with human-centred approaches to cyber-attacks targeting the AI component of smart devices, hereon AI-enabled devices, which exploit “inherent limitations in the underlying AI algorithms” (Comiter, 2019). Such cyber-attacks can affect the *integrity* of the system through poisoning attacks that corrupt the AI dataset or model during its training, or evasion attacks that provide input that leads to incorrect AI output (Comiter, 2019; Pitropakis et al., 2019). Users, however, remain largely unaware of

the range of possible cyber-attacks on AI-enabled devices – and why these might pose an issue for them in the first place (Bouwmeester et al., 2021; Meneghello et al., 2019; Spero and Biddle, 2019). In recognition of the intrinsic ambiguity introduced by AI and the lack of predictability in how it behaves, it has been shown that AI can remain opaque to users (e.g. Peters, 2023), highlighting the possible interpretive challenges users may face whilst making sense of cyber-attacks performed on the AI.

Against this socio-technical landscape, some countries, such as the UK, have set national policies on cybersecurity that expect from individual citizens an active role in managing their own security risks and developing personal strategies to protect themselves (Turner et al., 2022). In response, previous research has tended to focus either on developing technical solutions that increase AI-enabled device security without user involvement (Chalhoub et al., 2021; Hammi et al., 2022; Rahman et al., 2021; Rostami et al., 2022; Slupska et al., 2021), or equipping users with preventative actions that can minimise cybersecurity risks (Turner et al., 2021). To date there has been little research to investigate how the distinctive character of AI affects users' understanding of cyber-attacks on their AI-enabled devices and furthermore, how we might design interventions to support the active role expected from users in mitigating cyber-attacks in the context of this emerging technology.

Our research seeks to address this gap by introducing and evaluating a new cybersecurity intervention designed to support home users of AI-enabled devices to identify cyber-attacks on the AI and take mitigative actions in the aftermath (i.e. to return the device to a secure state). To advance the understanding of the understudied cybersecurity dimension of domestic AI applications, we use smart heating as a case study allowing us to draw transferable insights to how future cybersecurity interventions can be designed. We focus on integrity attacks which have the potential to be perceived by users and are also the most prevalent in the smart home cybersecurity literature (Heartfield et al., 2018). In the winter of 2023, we invited ten UK households (18 participants) to adopt a bespoke smart heating system called 'Squid' for a period of seven weeks, during which it suffered unannounced cyber-attacks of three different types. To better perceive and diagnose these attacks, participants had access to a cybersecurity intervention consisting of: i) multimodal indicators, including visual *Explainable AI features*, incorporated in Squid's application that brought attention to a/typical AI behaviour, and ii) a separate *assisted remediation tool* (helper tool) that supported the interpretation of these features and proposed *remediation actions* tailored for each cyber-attack type. Through a rigorous mixed methods approach involving technology interaction logs, semi-structured interviews, and self-reports with diaries, we address three research questions: Which types of AI integrity attacks are easiest to spot and what indicators contribute to this? (RQ1) How do users interact with Squid's features and its assisted remediation tool to mitigate an integrity attack? (RQ2) How do users negotiate cybersecurity in the context of AI? (RQ3).

2 Background

2.1 The Role of the User in Cybersecurity

Within the cybersecurity literature, there has been ongoing debate on whether, and how, users should be involved in mitigating cybersecurity risks and attacks. However, research addressing the role of the user in this domain remains relatively scant. A common narrative situates the user as part of the problem (Jeong et al., 2021; Still, 2016), with the prevalent 'secure by design' approach removing the user as an active agent entirely (Rostami et al.,

2022). Where users' role has been considered, the focus often falls on pre-emptive measures, such as the implementation of password managers (Chalhoub et al., 2021; Hammi et al., 2022; Slupska et al., 2021; Turner et al., 2021).

In envisioning what it means for users to adopt a more active role, Frik et al. (2019) explain how taking personal responsibility for cybersecurity "requires understanding of communicated risks, the opportunity to act, and to know how to act". In reference to smart devices, Meneghello et al. (2019) also highlight the importance of cybersecurity education by showing how its absence leads to the neglect of even the most simple and important security measures amongst users, such as changing the default password of their smart devices. This can in turn allow attackers to exploit the device more easily as part of a botnet attack, without the user being aware of a security breach. This educational lens has underpinned bespoke initiatives (whether online or within the home) designed to support users to make more informed choices and understand the trade-offs between device functionality and security (Benton et al., 2023; Rostami et al., 2022). However, in achieving this vision, previous work has also pointed to associated challenges such as users' apathy or lack of motivation to engage with potential cybersecurity threats (Zimmermann and Renaud, 2019), as well as inappropriate responses to attacks – for instance, the attempt to transfer the same cybersecurity strategies across different types of devices (Bouwmeester et al., 2021; Zeng et al., 2017).

Offering a different perspective, other research has underscored the importance of understanding cybersecurity as a social process (Dourish and Anderson, 2006). Recent work has shed light on the variety of 'non-technical' protective actions people can take, such as Warford et al. (2022), who identified categories of protective practices that included *social strategies* (e.g., drawing on help from friends, family or organisations) and *distancing behaviours* (e.g., limiting what is shared online and reducing/ discontinuing the usage of concerning devices). To our current interest, the home – and its social, material, and technological qualities – can introduce further considerations in how users 'do' cybersecurity as a technical and social endeavour, and thus shape the roles they are realistically willing, or able, to take. For instance, the temporality of the home, such as daily routines of caring for children or getting ready for work (Pink et al., 2017), tends to constrain how and when people engage with smart heating technologies (Jensen et al., 2018a; Vasalou et al., 2024), a finding that likely extends to cybersecurity. Furthermore, previous work indicates that individuals' diverse levels of engagement and expertise with AI-enabled devices can lead to power imbalances intersecting with other categories such as gender, age and ability. The most tech-savvy member typically assumes control of the device, which by extension potentially limits other household members' agency at home (Ehrenberg and Keinonen, 2021; Nicholls et al., 2020). The ways in which occupants interact with each other, with technology and with the physical space of the home in everyday life can therefore shape their approaches to cybersecurity measures and thus attacks may be experienced in different ways (Benton et al., 2023).

2.2 Cyber-attacks in AI-enabled Devices

As will be discussed, the cybersecurity intervention proposed in this research aimed to support user understanding and skills to mitigate cyber-attacks on the AI. Cybersecurity research has begun to explore whether users are likely to discern specific indicators that signal a cyber-attack on an AI-enabled device, including in the AI itself (Bouwmeester et al., 2021; Chen et al., 2021; Heartfield and Loukas, 2018a). This body of research suggests that cyber-attacks on the AI are sometimes, but not always, perceived. Devices tend to have limited means to signal abnormalities, resulting in cyber-attacks being indicated by nuanced,

or seemingly unimportant, changes in the AI-enabled device, which are often ignored (Comiter, 2019; Kuzlu et al., 2021). Many AI-enabled devices, such as popular voice assistants, do not incorporate a screen and can only communicate persistent feedback through simple operations, such as turning on the device light, which limits the granularity of any potential indicators (Rostami et al., 2022). This could partially explain why users find it challenging to interpret the device behaviour, whilst struggling to differentiate between system errors in AI-enabled devices and an actual cyber-attack (Huijts et al., 2023; Rostami et al., 2022). Moreover, in cases where users have received notifications about an attack, as Rodriguez et al. (2022) have found, these tended to be initially ignored. In this study, just over half of participants were notified more than once, with 11% requiring 10+ notifications. Users' slow response resulted in the overall cyber-attack lasting longer than expected, highlighting the importance of ensuring that notifications are quickly noticed and acted upon (Rodriguez et al. 2022).

In addition to the difficulties involved in identifying cyber-attacks, users face several challenges in attempting to remediate the situation, such as the lack of security guidance in device manuals and the generic advice they receive from their Internet Service Provider (ISP). Although this is partially due to data protection regulation restrictions on identifying specific devices, Bouwmeester et al. (2021) show that even when the ISP has provided users with steps to deal with infected devices, most users failed to complete them (e.g., changing the password on the device/router; disconnecting device; restarting/resetting device and/or resetting the router). Furthermore, there are new forms of persistent threat that cannot simply be resolved through standard approaches such as restarting/resetting devices and require more complex actions from users (Rodriguez et al., 2022). A key challenge thus stems from the overall limited technical tools that can practically support users to perform the remediation (Rodriguez et al., 2022; Zimmermann and Renaud, 2019).

Against this complex landscape, the available research concerned with designing and understanding how users can remediate their AI-enabled devices has not yet considered how this extends to attacks on the AI component, a gap we address in our research. Given the current lack of AI-specific guidelines, we draw on broader literature concerned with the cybersecurity of connected devices. Past work suggests that users will have a better chance at identifying and recovering from a security breach if: (i) manufacturers support user awareness of major threats and remediation strategies, and (ii) devices provide more information about their operations (Turner et al., 2022; Zeng et al., 2017). This can include indicators that can help users identify suspicious behaviour, such as login/settings change logs and alerts (Bouwmeester et al., 2021; Spero and Biddle, 2019), as well as behaviours that may be commonly mistaken as hacks (Rostami et al., 2022). Moreover, it has been suggested that smart devices, more broadly, could offer some form of *assisted* remediation, such as a knowledge base providing the user with a set of actions and tasks (Bouwmeester et al., 2021; Rostami et al., 2022; Turner et al., 2022).

This body of literature informs the initial conceptual underpinnings of this paper and guides our efforts to design an AI-enabled technology intended to support cybersecurity diagnosis and remediation. Table 1 summarises this work in the form of key principles.

Table 1. A set of principles for designing new AI-enabled devices to support cybersecurity diagnosis and remediation and key challenges resulting from the AI component of these devices

Cybersecurity design principles for smart technology	**Related literature**
Smart device designers should provide guidance about how to identify potential cyber-attacks in their devices, including 'normal' behaviours that may be commonly mistaken as attacks.	(Bouwmeester et al., 2021; Huijts et al., 2023; Rostami et al., 2022)

Smart devices should make explicit the actions they are performing, as part of their normal behaviour, to allow the user to easily monitor these.	(Turner et al., 2022; Zeng et al., 2017)
Smart device manufacturers should share a set of clear perceivable cyber-attack indicators.	(Comiter, 2019; Kuzlu et al., 2021)
Smart devices should incorporate some form of assisted remediation (e.g., helper tool) for common types of attack.	(Bouwmeester et al., 2021; Rostami et al., 2022; Turner et al., 2022).
Smart devices should provide clear feedback on any remediation actions to help users determine when an attack has been resolved.	(Bouwmeester et al., 2021; Rostami et al., 2022)

2.3 Motivation and research questions

In summary, this paper is motivated by the relevance of cyber-attacks on AI-enabled devices that are becoming increasingly embedded in the physical realm, particularly in our homes (Loukas, 2015), and the onus for the user to resolve them. We respond to the limited research directly engaging with the particular challenges involved in cyber-attacks on AI-enabled devices, where the system behaviour is already often perceived as unpredictable and incomprehensible, even before an attack has occurred. Despite some guidance on how to support users with connected devices, we have identified a research gap around how smart devices with AI should be designed to enable users to successfully understand and address cyber-attacks, for example through specific attack indicators. We argue that shedding light on this issue can inform how we might design future AI-enabled devices to include indicators that are most likely to be perceived (e.g. a light showing a microphone is in use) or how devices could be engineered to include physical safeguards (e.g. a smart camera lens shutter).

To address this gap, we designed a new cybersecurity intervention around an AI-enabled technology for domestic use – a bespoke smart heating system we called "Squid". Using Squid as an exemplar of an AI-enabled device we seek to contextualise cybersecurity principles to support users to mitigate integrity attacks performed on the AI component. Thus, following an introduction to Squid, in Section 3, we draw on the principles reported in Table 1 to present the intervention consisting of: (i) indicators (e.g., Explainable AI features, physical device outputs) designed to communicate typical and atypical AI behaviour; (ii) remediation actions; (iii) an assisted remediation tool designed to scaffold the use of these features to navigate the nuances of cyber-attacks on the AI. Our overarching research aim is to understand if and how this new cybersecurity intervention supports users to identify, diagnose and remediate integrity cyber-attacks on AI in the home context. Taking an exploratory approach, that pays attention to how everyday life factors can interact with new interventions in this context, we ask the following research questions:

- **RQ1**: Which types of AI integrity attacks are easiest to spot and what indicators contribute to this?
- **RQ2**: How do home users interact with Squid's features and its assisted remediation tool to mitigate an integrity attack?
- **RQ3**: How do home users make sense of cybersecurity in the context of AI?

3 SQUID: A SMART HOME TECHNOLOGY FOR HEATING

Squid is a smart technology we developed that uses AI to regulate heating. It was inspired from previous research by Alan and colleagues (Alan et al., 2016), who used AI to develop a smart thermostat supporting home users to adjust their heating based on dynamic tariffs,

which are increasingly introduced into the context of domestic energy use. Drawing from this research, Squid's AI is premised on the assumption that people have sensory preferences that inform their ideal room temperature, but also have a preference as to how much they are willing to pay for energy. Thus, Squid's automation allows for frequently adjusting the temperature settings throughout the day in response to the varying price conditions, removing the need for time-consuming manual interventions.

For the purposes of this research, Squid was designed to operate in one room regulating the heating of a single radiator. It included three components (see Figure 1) – two were physical components from the commercially-available Netatmo smart radiator valve kit, and the third, a web app, developed as part of the research:

- The physical **smart valve** is installed on a radiator. It is fitted to the pipework at the bottom of the radiator replacing any existing thermostatic radiator valve (TRV). The valve includes an embedded thermostat that regulates the flow of hot water into the radiator by opening or closing the valve. It contains a physical display that shows the current temperature as measured by the valve, as well as the target temperature set by Squid's web app for that time.
- The **relay** is plugged to an electrical socket, and connects both to the internet and the **cube**, which is the Netatmo physical interface with users. Both include a reading of the temperature at the valve and the ability to change the thermostat. The functions of the cube were hidden to our participants by a sticker, and they were only told that it manages connectivity and should be placed in the same room at the valve. A QR code placed in both the cube and the relay lead users to the *helper tool* from their mobile phones.
- The Squid **web application** was developed to allow AI control of the smart valve, which the user trains using the app.

Smart valve | Web application | Cube | Relay

Figure 1. Squid ecosystem

3.1 AI Model

Squid uses a dynamic pricing model in which energy prices fluctuate every 30 minutes, proposing an economic way of reasoning about energy. With the automation, a household acting economically rationally could save money by lowering the target temperature when the price is high and increase it when the price is low to sustain their thermal comfort. Our participants were initially required to make deliberate choices on how to balance the energy price at a given time with the temperature that matches their thermal comfort preference. Based on their temperature inputs at given price points, Squid's algorithm extracts each user's sensitivity to price and calculates their preferred temperature for each of five time slots in a day (see 3.2 for more details). We note here that as energy prices are not publicly available for more than a day in advance, historical prices were used, reflected in the 2019

dataset published by Octopus Energy, which was one of the first energy providers to introduce dynamic pricing to households.

Bayesian linear regression was used with one input feature (price) and one target feature (target temperature). This choice was made for two reasons. Firstly, Bayesian linear regression is an example of *interpretable* (or glass box) machine learning, meaning that its model parameters can be inspected. Restricting the model to two features makes it a simple and comprehensible instance of a glass box machine learning model[1]. Secondly, Bayesian regression models can start being used with a predefined default model that can be updated immediately after each input, which makes it usable from day one, as well as after any AI model reset that the user may request. These choices allow us to explore our new approach to cybersecurity in a simple scenario providing opportunities to expose the AI models to users through visualisations and other Explainable AI methods. We note here that the price sensitivity parameter was also presented to users according to one of six categories (negative, very low, low, moderate, high, very high) determined through our own normalisation procedure. By default, Squid was predefined to exhibit a preferred temperature of 22ºC and moderate price sensitivity, which adapts gradually as the users begin to change their target temperature. This is the default the state the model returns to when the user resets the profile.

3.2 Squid's features

Squid incorporates a **schedule** page that presents five pre-populated heating profiles and their time slots, which can be adjusted for any day of the week (see Figure 2). Through its **side panel**, which is always visible, it displays the current temperature, as measured by the smart valve's thermostat, and the target temperature determined by the AI. Using the temperature dial on the side panel, the user updates the temperature allowing the algorithm to learn. The side panel also presents current dynamic tariff, i.e., the energy prices for the 30-minute slot at that time, alongside statistics for daily/weekly/monthly prices. The prices are presented in the sector's standard format of pricing in pence per kilowatt hour (p/kWh), which is the same used in regular energy bills. In addition to this key information, the side panel displays when the smart valve is open, which heating profile is active and offers a button to switch Squid to manual mode, disabling the AI.

[1] A Bayesian linear regression model forms a multivariate normal distribution parameterised by a mean vector and covariance matrix, which can be assigned intuitive meaning. Specifically, by using only these two parameters, the model becomes a bivariate normal distribution with the slope being exposed to the users as the "*price sensitivity"* and the y-intercept as their "*preferred temperature if energy were free*".

Figure 2. Squid heating profile schedule and side panel

Squid's features were designed to offer customisation and visualisation of how the algorithm works to inform and automate heating. Drawing from recent research in Explainable AI (Abdul et al., 2018), the same visualisations were used to provide indicators in the event of a cyber-attack (see 4.2). Development of Squid followed the XAI principle of interpretability-by-design where the AI feature was implemented using a simple glass-box machine learning model with the model subsequently exposed to users via the web interface, including text, image and interactive visualisations. We follow the widely accepted definition of interpretable AI as provided by Murdoch et al. (2019) to be "the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model", where it is clarified by the same authors that "we view knowledge as being relevant if it provides insight for a particular audience into a chosen problem".

- Under the schedule tab, users can click on any heating profile. This triggers the AI **temperature schedule visualisation** (see Figure 3a) that displays how the AI is expected to change the smart valve's target temperature, whilst the price changes during the selected time slot. According to the default model, when the price (in orange) goes up, the target temperature (in teal) decreases.
- Under the profiles tab, users can find a summary of the AI model for each heating profile. When the **gauge visualisation** (see Figure 3b) on the top part of the page points to the right red section, it indicates that the household is very sensitive to price (i.e. the target temperature will decrease sharply as price increases). In contrast, when the gauge points to the left red section, this indicates that the household has negative price sensitivity (i.e. the target temperature will increase as price increases). Underneath the gauge, there is an **AI summary visualisation** (see Figure 3b) that summarises the current AI model's target temperature prediction with respect to price[2].
- Within the profiles tab, clicking on "**want to know more about this profile**?" takes the user to an additional page (see Figure 3c), which displays (clockwise from the top left) (i) the set of prior user inputs as data points on a temperature vs. price chart; (ii) the mean vector and 99% confidence interval of the Bayesian linear regression model after those prior user inputs; (iii) the same chart as in Figure 3b but with the addition of the 99% confidence interval; and (iv) the same chart as in Figure 3a but with prices taken from a selected day rather than a selected time slot. All four charts can be controlled using the

[2] Technically this chart shows the *mean* predicted by the Bayesian linear regression model, so the model's uncertainty is not represented, and it appears as is typical for simple linear regression.

buttons located at the bottom of the page. Selecting the forward and next buttons labelled *Inputs* allows users to inspect any prior AI model up to the last profile reset. Selecting the forward and next buttons labelled *Day* allows users to inspect the predicted target temperature schedule for that day if the profile were active for the entire day.

- Finally, under the **notifications** tab (see Figure 3d), the user can view their own temperature inputs and the decisions the AI has taken for each thirty-minute slot based on the price applying to that slot. Specifically, for each AI decision, the notification provides a summary of the decision that includes the preferred and target temperatures (in degrees °C), the price sensitivity (displayed as "very low" to "very high", mapping to the six sections of the gauge), and the energy price that was considered.

Figure 3. Squid. (a) AI temperature schedule visualisation ; (b) AI summary and gauge visualisation. (c) What to know more about your profiles page. (d) Notifications log

4 CYBERSECURITY IN THE CONTEXT OF SQUID

4.1 Integrity cyber-attack types presenting in Squid

Three *integrity* cyber-attacks were possible to trigger in Squid, which interfered with the functioning of its AI, specifically with the aim to emulate lowering the target temperatures. We chose integrity attacks because they lead to impact that is directly observable by the participants. In contrast, an availability attack, that makes the system inaccessible, would have been indistinguishable to normal service downtime, and a privacy attack would have left no trace that would be observed by a non-expert user. Following the standard classification of AI attacks on poisoning and evasion (Pitropakis et al., 2019), we chose one poisoning integrity attack where it is the model that is targeted, and one evasion integrity attack where it is the input data that is targeted. For the poisoning attack, we also developed a variation that would be harder to detect. We kept the total number of attack types to three to stay within the timeframe constraints of the field study, while still covering both categories of AI attacks and allowing for attacks of different complexity.

The simple **AI poisoning attack**, hereon *Attack 1 (A1)*, involved the input of multiple false temperature data with unrealistically low values (between 7.5 and 10C) over a period of 20 minutes (two entries per minute), which heavily modified the AI model towards an extremely low preferred temperature and pushed the sensitivity to price outside its usual range. These target temperature entries became part of the user's active heating profile such that Squid began to heat the radiator according to this new (poisoned) data. The complex **AI poisoning attack,** or *Attack 2 (A2)* was a variation of A1. It involved the input of false temperature data reducing the target temperatures, which were entered into the system over an hour, becoming part of user's active heating profile. However, this attack also deleted the logs associated with the false temperature data entries. *Attack 3* (A3) was an **evasion attack** that involved the manipulation of the price data that was used by the Squid AI model to regulate the temperature, mimicking an attack to Squid's headquarters. Squid was thus using the wrong price data and subsequently calculated the temperature incorrectly. The attack was restricted to a period of five hours, after which the price data was restored to its original, correct state.

A1 and A2 had an enduring effect on one of Squid's heating profiles, which continued to affect the heating in the room unless the user acted to address the attack and were thus more persistent compared to A3 which lasted for a fixed period. The sensory experience of A3 thus depended on household occupants being present in the home when it happened. Table 2 summarises the characteristics of the three attacks.

Table 2. Attack summary

Attacks	Indicators	Mitigation	Persistence
Attack 1 (A1)	• Low temperature values in the log registered during the attack, and after the attack (when the heating profile is active) – *visual (Fig 3d)* • Excess logs added during the attack – *visual (Fig 3d)* • Gauge pointed to the red section – *visual (Fig 3b)* • Valve buzzes more frequently during the attack – *auditory* • Room gets colder - *sensory*	Reset profile	Heating profile affected until action taken
Attack 2 (A2)	• Low temperature values in the log registered during the attack, and after the attack (when the heating profile is active) – *visual (Fig 3d)* • Gauge pointed to the red section – *visual (Fig 3b)* • Valve buzzes more frequently during the attack – *auditory* • Room gets colder - *sensory*	Reset profile	Heating profile affected until action taken
Attack 3 (A3)	• High price values and low temperature values in the log registered during the attack – *visual (Fig 3d)* • AI interaction graph allowing users to view the model's behaviour historically, on the day and its future predictions – *visual (Fig 3c)* • Room gets colder while attack is active - *sensory*	Set to manual	Heating profile is affected for the 5-hour duration of the attack

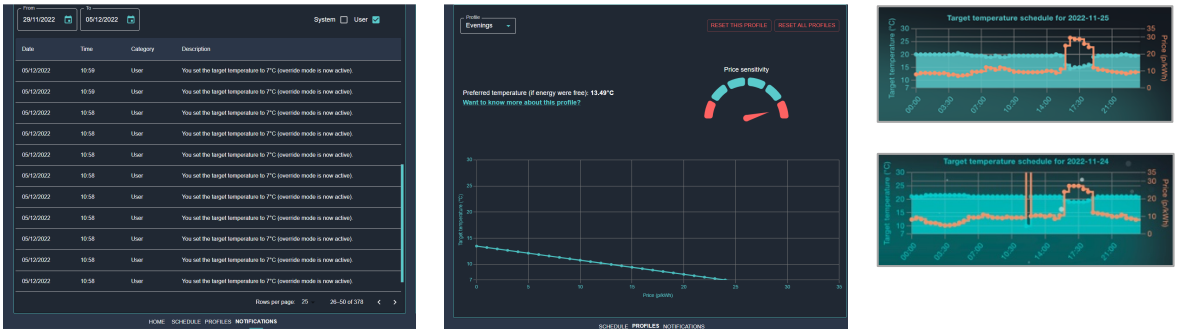
4.2 Cybersecurity indicators, diagnosis, and actions

A key challenge in supporting users to mitigate a cyber-attack are the range of possible attack types that could affect technology in different ways with each requiring a possibly different response. Informed by previous research highlighting the importance of incorporating clear

indicators and meaningful actions in the design of smart technology (see design principles under 2.2), Squid and its explainable AI features were used to support users to notice, diagnose, and act on each attack.

As Table 2 illustrates, with regards to indicators, during **A1**, it was possible to observe the new entries introduced by the attack within Squid's *notification log*. The introduction of extremely low temperatures also affected Squid's *gauge* which now pointed to a red section and thus indicated price sensitivity that is outside the reasonable range. **A2** was expressed within Squid in an identical way except for the *notification log*. Looking at the logs, it was possible to observe unusually low temperatures, but it was harder to spot the attack since the excess false inputs had been deleted. Finally, during **A3**, since the AI model was not affected, the gauge remained in its usual position. If the user scrolled through the *notification logs*, they might notice the unusually high prices associated with the attack. Moreover, in enabling the user to interact with the historical models and its future predictions, it was possible to observe the price peak introduced by this attack in the 'want to know more about your profiles' chart. Figure 4 presents how these indicators appeared to Squid's users. We note, that in addition to the indicators present in web application, during A1 and A2, the *smart valve* was activated to open/close more frequently than the change of the energy tariff, introducing a recurring buzzing sound for the duration of the attack.

In response to these attacks, users were provided with specific actions. A1 and A2 were mitigated by resetting the poisoned heating profile and thus reverting to the AI default model using a button on the top of the profile page. This removed the new data entries generated by the attack. However, it also introduced the need to re-engage with the heating profile and train it to reflect users' preferred temperature. Within A3, the action was to switch Squid to manual mode to pause the impact of the attack until it ended.



Notifications-logs (a) *Gauge (b)* *AI interaction graph (c)*

Figure 4. (a) extra user logs with low temps registered in the heating profile, (b) gauge is irregular (red); (c) AI interaction graph shows irregularity in price during the time of the attack compared to other days

4.3 Helper tool

Reflecting the need for assisted remediation outlined in Section 2.2 we recognised the need for a user-facing cybersecurity intervention that is simple to use, while also designed to support users in specific and variable circumstances. Against these considerations, we designed a 'helper tool' to scaffold the user to interpret the information within Squid. Underpinning the tool, was a flowchart of diagnostic questions (between 4-5) that resulted in a proposed action appropriate to that attack, akin to the troubleshooting guides found in the manuals of typical domestic appliances. The approach was inspired by the diagnostic flowcharts used for fault detection in industry as well as tree-based modelling of cyber risks and faults (Nagaraju et al., 2017) and tree-based detection (Vuong et al., 2015) of threats to cyber-physical systems like vehicles.

Turning to the design of the helper tool, each question-check asked the user to evaluate how Squid was functioning and inspect the Explainable AI. Each question-check presented a visual to ensure it was clear to the user what to check in Squid's interface or ecosystem. Depending on their replies, users took different paths that resulted in a binary assessment on whether there was a cyber-attack. When an attack was detected, the user was presented with a 1-minute video that displayed the attack type, how it affected Squid and the appropriate action. Figure 5 presents the question-check flow for each attack. As explained under the Methodology (Section 5), all participants received training around how to use the cybersecurity intervention, which included the helper tool.

Note that a cyber-attack on the AI of a smart heating application means that either the user's device or the Squid headquarters has been compromised. Had a helper tool component been incorporated directly into Squid, it would also be subjected to manipulation during an attack. To address this, the helper tool was designed as a separate application.

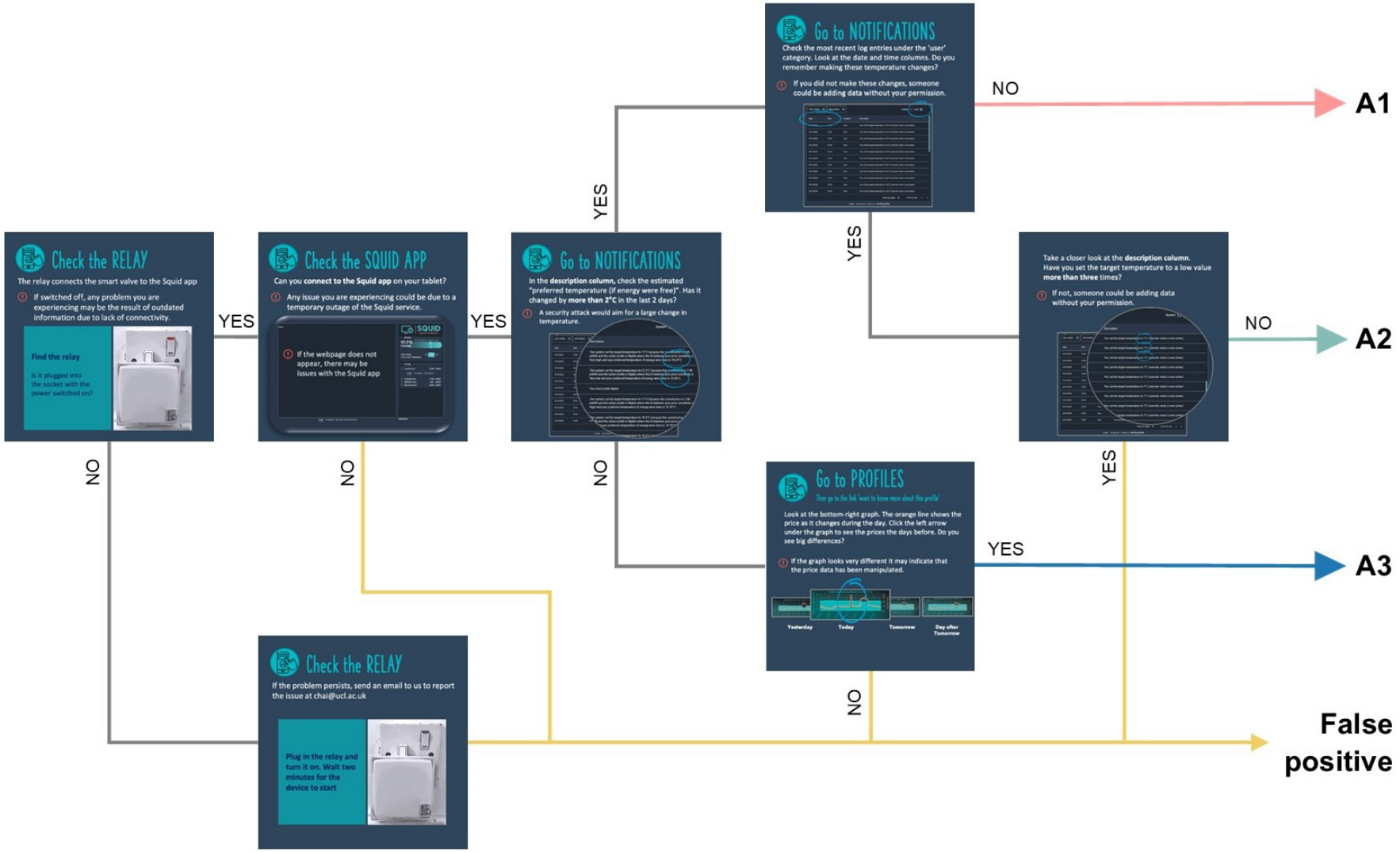


Figure 5. Helper tool questions for the three attacks and an example false alarm

5 Methodology

5.1 Recruitment

The research took place between January and March 2023 with households in England. It is important to note that the study coincided with a surge in energy prices and increases in the cost of living more broadly (Guardian, 2023). While this circumstance underscored the relevance of Squid's AI, the team was mindful that many households faced added vulnerabilities during winter. Thus, in addition to recruiting households who could commit to engaging with the study and had the level of digital and data literacy needed to use Squid, we ensured the households involved were not vulnerable to the cost of living or sudden shifts in home temperature.

The study was advertised through community centres, social media, and snowball sampling through our existing networks. As an incentive, the project offered compensation in vouchers that amounted to £120. During the period of the cyber-attacks, participants were advised that each attack diagnosed amounted to £10 (earning them a maximum of £60 for a total of six attacks). This reward was introduced to incentivise the continuous engagement required to interact with the cyber-attacks (also see 5.3.2, a methodological decision which we reflect on in the discussion).

Households were recruited based on two criteria. First, we wished to encompass a range of living arrangements and household compositions. Second, we wanted to identify participants with a mix of professional backgrounds. This sampling approach was deemed important for us to capture heterogenous heating practices and different ways of interacting with AI. Thirty-six households were initially identified/expressed interest, and eleven households signed up with one household dropping out mid-way, due to personal reasons. Thus, ten households (18 participants) took part and completed the study. Except for one household, all others were recruited through the research team's extended social and professional networks.

5.2 Participants

Within the above constraints we gathered a diverse cohort of participants in terms of professional background, household composition and living arrangements, although all of them in middle-class areas. Except for one participant, all others had earned a higher education degree. All participants were competent technology users, and seven of the households owned their home. From the ten households involved, four were in London and six in the Southwest of England. Of the 18 participants, four were born and raised in Britain. The remaining participants had immigrated to England from nine different countries (Turkey, Hong Kong, Taiwan, France, Germany, India, Spain, Slovakia, Greece, Brazil). As Table 3 indicates, most households consisted of nuclear families and couples. Participants' dwellings were balanced between single occupancy houses and flats. Only one couple lived in shared accommodation. All participants used technology as part of their everyday lives and reported being confident in using technology to participate in professional and social activities. Except for one household (Sam & Mara), the rest used at least one smart device. Two households (Isaad; Barış & Maya) owned between three-four smart devices and considered themselves technology enthusiasts.

From the eight households that had two adult occupants, one of them led on the use of Squid with the other playing a supportive role. This supportive role consisted of either using

Squid occasionally when the lead user was unavailable, or prompting the lead user to engage with Squid e.g., when sensing the room was cold.

Table 3. Participating households (participants' names have been pseudonymised). The lead user is indicated with an *. Participants who occasionally used Squid are indicated with a ^

Participants and their professions	Household composition	Dwelling	Smart devices
*Simone, primary school teacher (French) ^Theo, medical writer (British)	Parents, two children	Semi-detached house	Smart watch, smart meter
*Isaad, head of service delivery (British/Moroccan)	Father, two children	Detached house	Smart meter, personal assistant, smart doorbell
^Wiola, academic university staff in health (German) *Naadir, town planning consultant (Indian)	Parents, two children	Semi-detached house	Smart watch, smart meter, personal assistant
*Ernesto, research manager (Spanish) ^Klara, medical trainee (Slovakian)	Parents, one child	Semi-detached house	Smart meter
*Carrie, Communications manager (British)	Single occupant	Three-bedroom terraced house	Smart meter
^Maya, post-graduate student (Turkish) *Barış, Civil engineer (Turkish)	Couple	One-bedroom flat	Smart watch, smart meter, smart doorbell
^Sam, event worker (British/Greek) *Mara, academic university staff in education (Greek)	Mother and adult son	Two-bedroom flat	None reported
*Kevin, marketing (Hong Kong) Ria, sales (Taiwanese)	Couple flat sharing with three flat mates	Four-bedroom flat share	Smart vacuum cleaner
Açelya, CEO of a start-up in education (Turkish) *Ender, CEO of a start-up in accessibility (Turkish)	Couple	Two-bedroom flat	Smart watch
*Debora, head of R&D (Brazilian) ^Aris, academic university staff in Microbiology (Greek)	Parents, two children	Semi-detached house	Personal assistant

5.3 Field study procedure

The study received ethical approval from all participating universities' ethics committee. Following initial contact, participants were invited to an introductory conversation about the research. The purpose of this conversation was to share the aims of the research and ensure participants were aware of and able to commit to the proposed tasks. Once they registered interest, an information sheet and consent form were provided. All adults in each household completed the informed consent and participated in the data collection. Table 4 introduces the field study research design which we describe in detail in this section.

Table 4. Summary of field study

	Phase 1: Familiarisation with Squid (3 weeks)	Phase 2: Using the cybersecurity intervention (4 weeks)
Engagement with household	• One house visit (install technology) • Training on how to use Squid to heat the focal room • Weekly online check-ins	• Two house visits (uninstall technology) • Training on how to use Squid to use the cybersecurity intervention • Weekly online check-ins
Data collection	• One interview • Weekly fieldnotes • Squid interaction logs	• Two interviews • Weekly fieldnotes • Diaries • Squid interaction logs, helper tool logs

5.3.1 Phase 1: Familiarisation with Squid

The study started between the second and third week of January 2023 for all households. The research team visited each household to install the smart valve. To ensure the smart valve would have an impact on how people experienced their heating, we asked participants to select a radiator in a room they frequently used. Since the technology set-up consisted of a

single smart valve, we wished to install the smart valve in a room that had one radiator. If there was a second radiator, we consulted participants if they were happy to switch it off for the duration of the study, which two households did. Participants were given an 8-inch tablet hosting only the Squid app. This ensured that all households experienced Squid at the same screen resolution. Following the installation, a semi-structured **entry interview** explored participants' routines and everyday experiences with technology in the home as they related to cybersecurity and heating. This is reported in (blind for review).

Participants then received a 30-minute training session on how to use Squid, with a focus on supporting households' understanding of energy tariffs/dynamic pricing, the AI algorithm, and Squid's explainable AI features. Specifically, through hands-on interactions with Squid, participants used the app's features, learned how to set up profiles, set their temperature by consulting the current price, evaluate the chart relations and log content. To ensure participants' understanding of Squid, the same topics were reinforced through four 1-minute info videos sent to the households over the course of two weeks via email (two videos per week).

We aimed for each household to use Squid for three weeks during the familiarisation period. This was possible except for one instance (Simone & Theo) where we had to limit the period to two weeks due to the household's scheduling constraints. Participants were encouraged to calibrate the AI during this period and use the features introduced in the training as they saw fit. To maintain their participation and ensure there were no technical issues, the research team held 15-minute online check-ins with each household weekly. The check-ins were held with one household member to maintain a flexible approach. Thus, through the check-ins and the training, our aim was to support participants to understand Squid's design rationale and embed it in their heating practices.

5.3.2 Phase 2: Experiencing and mitigating cyber attacks

Following the familiarisation phase, we visited again the households and trained them in the cybersecurity intervention. The 30-minute training aimed to support participants in using the helper tool to diagnose and adequately act upon an attack on the AI. A live cyber-attack was performed on Squid, in front of them, allowing us to contextualise the training to the real experience of an attack. The training focused on A1 as an exemplar (detailed in 4.1), while we ensured participants were aware that other attacks could manifest in different ways.

Thus, using the helper tool during the training, participants were supported to recognise the cyber-attack indicators and use Squid to diagnose, as well as mitigate, the attack as recommended by the helper tool. We paid attention to supporting participants to associate the questions posed by the helper tool to Squid's explainable AI features. The helper tool was also used as if the attack was retroactive, reflecting that cyber-attacks could occur when participants were not at home. Since A1 required the resetting of an infected profile, participants were guided on how to accurately identify which heating profile had been affected (through the notifications-log page), reset it (through the profiles page), and retrain it by using the side panel. Table 5 summarises the training content and tasks the participants engaged in.

This second research phase lasted four weeks for all households. During this period, households experienced two simulated cyber-attacks for each attack type (six attacks in total), usually timed to occur when participants reported being most at home. We note that the order of the attack types was randomised within certain constraints. These included a technical requirement for A3 to happen for all households at the same time, and knock-on implications of having different start dates for each household as we strived to maintain at least one day between each A1/2 attacks.

Table 5. Training content and approach

Training aim	Interaction with Squid	Training activity
Learn how to access the helper tool	n.a	Opening the helper tool in their mobile phone
Learn how to notice a *live* attack	n.a	Use senses to detect the smart valve buzzing and the colder room temperature
Learn how to notice a *retroactive* attack	Access the profile page	Perform a check of the gauge
Learn how to interpret the helper tool questions	Features as suggested by the helper tool	Perform checks using Squid's features
Learn how to reset the infected profile	Notifications page and Profile page	Identify the name of the infected profile from the log and select the correct profile to reset under the profiles page
Learn to retrain the profile	Side panel	Use the side panel to input the preferred temperature

5.4 Data collection

At the start of phase 2, a **semi-structured interview** was carried out with all household members to get a baseline understanding of how Squid was being used**.** The dynamics arising from participants' use of Squid offered valuable insights to our interpretation of how they later engaged with Squid's cybersecurity intervention. We also collected information about the times household members were most likely at home ensuring we could time the cyber-attacks to when they would be most noticed. The interview questions centred on (i) participants' understanding of Squid's AI and its relevance to how they made heating decisions; (ii) their understanding and use of Squid's features; (iii) the times of day they used Squid and (iv) who used Squid the most and why, if the household included more than one adult.

During phase 2, a range of data were collected to capture what participants did during the one-month period the cyber-attacks occurred. Due to the longer-term nature of the study, there were concerns around participants forgetting key events and actions, making it important to include methods that could prompt their memory in the final interview. We also wanted to ensure we could capture behaviour-related data alongside participants' reported experiences, which led us to follow a mixed method approach. This data included:

- **Fieldnotes** generated from the 15-minute weekly online check-ins that continued during phase 2 with each household.
- Online **diary** entries kept by participants after each perceived attack answering contextual prompts about it (*who was present, what they perceived, if the helper tool was intuitive, how they felt*).
- Logs from the **helper tool** recording when it was accessed, and which paths participants followed.
- Logs from **profile resets, changes to manual mode**, and **temperature inputs** added by users alongside their timestamps.

In addition to the data recorded during the one-month period, an **exit interview** was carried out. The interview started by exploring the indicators participants generally relied on to notice an attack, their responsiveness to remediating the attack, which features they used and why, and how the helper tool was used during the four-week period. Following these general questions, the researcher introduced two diary entries participants had created. The entries were selected to reflect different experiences with the attacks, including challenges participants had faced. Each diary entry was discussed with participants to ensure the nuances

and complexities of noticing and diagnosing the cyber-attacks were reflected in the collected data. Having discussed participants' actual experiences, the researchers presented the three attack types that had occurred with the aid of a visual sheet, the sensory, auditory, and visual indicators associated with each attack, and the actions required to mitigate it (as detailed in 4.2). This allowed participants to reflect on which integrity attacks they had experienced and ease-difficulty in doing so.

In total, 20 interviews were carried out across the two visits reported here, the pre-study interview lasted an average of 56 minutes, and the exit interview was an average of 63 minutes. All interviews were audio-recorded, anonymised and transcribed.

5.5 Data analysis

Thematic analysis was carried out on the exit household interview and the fieldnotes collected during the intervention. The first author, who had coordinated the data collection during the fieldwork and was familiar with the households, drew on the data to develop descriptive summaries for each household that documented their interaction with the intervention. The summaries were shared with the broader team of five researchers involved in the data collection to ensure contextual information was not missed. This step supported the second author, who had not been involved in the data collection, in the process of familiarising herself with the data whilst also ensuring that any relevant insights from the data collection process were recorded. Following this initial step, the same author carried out thematic analysis using NVivo. For RQs 1 and 2, a deductive approach was taken where we developed codes that aligned with the RQs. For RQ1, we coded the reasons participants offered for why specific integrity attack types were easy, or more challenging, to spot with reference to the indicators that helped them. For RQ2, the codes focused on households' reported interactions with Squid and the helper tool. Self-reports on what Squid features were used to diagnose/act during an attack were classified deductively by feature type. Following this, we also inductively analysed the same data using an open coding approach to identify households' interaction patterns, the reasons behind these patterns, and everyday factors reported that shaped how they used Squid. For RQ3, we followed an inductive coding approach on the same data, which explored the challenges participants encountered with identifying and remediating the attacks in the context of Squid's design. These initial codes were collaboratively discussed between the first two authors which led to clarifications and refinements of the codes leading to further sub-themes.

In addition to this thematic analysis, an analysis was carried out on the log data, which was visualised in Tableau. The first figure (Fig. 6) visualised whether participants spotted the different types of attacks over the one-month period, whilst also revealing when participants reported false positives (left column). It also showed how successful they were in reaching the correct resolution within the helper tool (middle column) and whether they followed the correct action to resolve each type of attack (right column). The second figure (Fig.7) was a temporal visualisation of households' interactions with Squid and with the helper tool over the course of the study, mapped to the time stamps of the cybersecurity attacks. Profile resets are presented with black bars and temperature changes with grey bars, interactions with the helper tool are indicated with coloured bars and cybersecurity attacks are presented with coloured triangles.

6 Findings

6.1 RQ1: Which types of AI integrity attacks are easiest to spot and what indicators contribute to this?

In the exit interview, participants reported A1 to be the easiest to identify, followed by A2. A3 was found to be the most challenging to spot. This trend is also supported by the log data presented in the left-hand column of Figure 6, which demonstrates that A1 was detected by all participants and A2 was also detected except for one instance, whereas A3 was identified by half the households with only three detecting both occurrences of the attack. In addition to this, half the households identified one, or two, false positive attacks. We now reflect on the salience of the indicators that participants perceived and relied on to infer there was an attack.

Participants offered several reasons to explain why A1 was the easiest to spot. The high number of available indicators (four) expressed in this attack increased the possibility of noticing it, which Debora explained "… *the more things that you can notice, the easier because you don't always notice them all. You notice one thing and then you go and investigate further.*" The notification log was an important source of insight as it allowed participants to detect irregular fluctuations in the target temperatures and made unexpected user entries immediately visible. Testifying to this, Sam explained "*Yes, because you see […] the big fluctuations, and a lot of it happening. So, you know something is wrong, something is going on, because you're not the one who is changing it all the time.*" In addition to this, seven participants relied on their senses when the room was cold, which was also reported for A2. This sensory focus is not surprising given the long-term impact of both attack types on a heating profile.

Despite having received training to use the helper tool only on A1, in all but one case, participants spotted A2 and therefore, successfully transferred much of their learning and experience to remediate an attack with less cues. Nonetheless, most participants were not able to necessarily (or immediately) distinguish between the two attack types when prompted. Moreover, when presented with the different indicators between attacks in the exit interview, participants reflected on the added challenges involved in detecting A2 in relation to A1. Whereas A2 was immediately visible through a single visual indicator, the gauge, the absence of excess logs in A2 added to the requirement to scrutinise the log content. Some participants who had detected both A2 attacks, speculated on future challenges they could face due to the relative subtlety of the indicators associated with this attack, such as being absent from home resulting in the attack being missed.

A similar pattern of findings was evidenced in A3, where the five participants who had noticed the attack and even reacted to it, struggled to point to indicators that would distinguish between attacks and still lacked an understanding of what had caused it. This was the case with Carrie who explained "*The fact that there could be a system attack on my user profile or on the system, I hadn't really thought about that big scale attack implication."* Moreover, as we will elaborate under RQ3, due to the unexpected nature of this attack, two further participants had noticed the attack unfolding, but dismissed the low temperatures triggered by A3 as normal behaviour of the AI algorithm explaining them away due to the high prices they observed. Besides the gauge not being triggered in this attack, we note that the relatively low rates of identifying A3 could have been due to its time-limited influence on the heating profile, thus highlighting the importance of householders' physical presence in the home at the time of the attack.

It is also noteworthy that half the households identified between one and two false positive attacks, which we describe under RQ3. In these instances, participants used the helper tool to disambiguate the event and confirm an attack had not occurred.

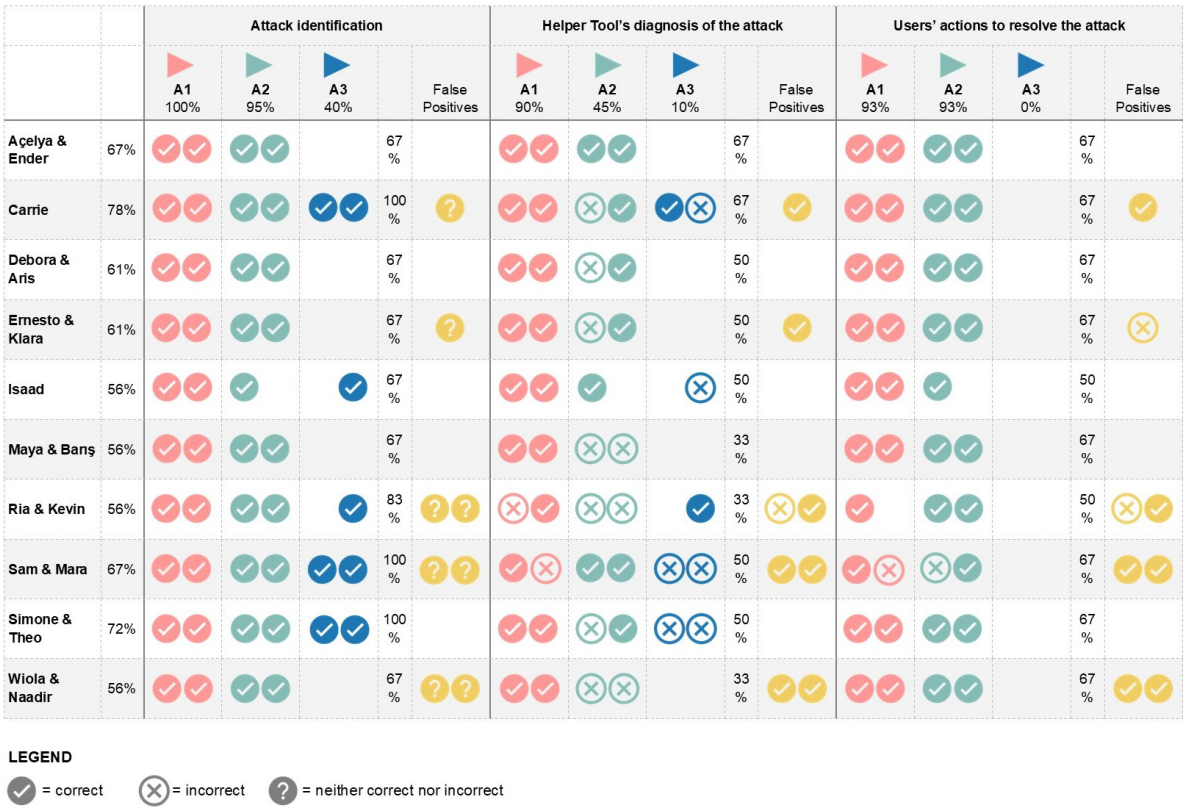


Figure 6. Overview of identified attacks (RQ1), use of helper tool to diagnose the attack and successful resolution of the attack (RQ2)

6.2 RQ2: How did users interact with Squid and its assisted remediation tool to mitigate an integrity attack?

6.2.1 Using Squid as a diagnostic tool

Before using the helper tool, all participants initially discerned by themselves whether there was an attack through the Squid web app to confirm its occurrence. Using the indicators reported under 6.1, their checks involved identifying if the current temperature or price was unusually high or low, examining the regularity of the current price sensitivity, or looking for temperature inputs they had not made. This was achieved in **different ways using Squid's features**, as we illustrate with the following examples. Barış primarily relied on the colour indicator of the gauge, subsequently spot-checking the notification logs to corroborate the presence of an attack. Simone checked the gauge and the shape of the slope under it to establish the presence of an attack, which she then verified by carefully reviewing the target temperatures set in the notification logs. Theo and Sam found the gauge challenging to interpret. Consequently, both participants relied on inspections of the notifications and Theo also used the side navigation to detect irregularities in the target temperature. Wiola looked at both gauge and notification logs, recognising that an attack may be visible in only one of these features. Across all testimonials the notifications-log emerged as the most significant visual indicator of an attack, albeit for some participants the logs being challenging to read.

Fifteen participants **proactively** searched for attacks by regularly (often daily) checking Squid's app. While the expectation of attacks was (unsurprisingly) a driving force for many, **the living situation, temporality, and environment of the home** required some to take a proactive approach to truly engage with the attacks. For instance, Carrie attributed her searching to being generally vigilant due to living alone. Sam's and Simone's proactive checks were motivated by not being in the focal room very often, with three other

participants, including Barış, taking a similar strategy due to not spending a lot of time at home: "*Because we are not regularly staying at home, we're generally outside, so we were checking the tablet once we come*". Ender observed that the heat retention in his home was generally poor, meaning that it was not possible to always rely on his senses and therefore it was necessary to proactively look out for the attacks. To cope with the time demands of a proactive approach, participants **created routines** to identify quiet moments when they could interact with Squid daily, e.g., after dinner time, when children were asleep, before work, or at designated times of the day.

Conversely, some were able to apply a **reactive** approach where interactions with Squid were sensorially triggered either by unusual patterns of the valve buzzing or the room's low temperature, both of which imply a more frequent presence in the focal room. Examples included Wiola whose attention was drawn to the buzzing sound while eating dinner: "*I think it was most, we were most likely to spot it when we sat here and ate and heard the clicks",* or Aris who noticed the difference in temperature between rooms: "*I noticed two (attacks) myself because there was a difference in the temperature. This room was cold and the other was hot. I figured too just by the temperature, not by checking actively what's going on".* Nonetheless, due to the variable ways in which the valve operated, or other competing noises in the house, the **buzzing was not always evident**, and when it was, as Klara noted, there was an effort involved to tune into its sound. Due to these reasons, compared to the low room temperature, the auditory indicator was not informative in four out of the ten households.

As the study progressed, and the distractions of everyday life took over, five participants **transitioned from a proactive approach to reacting to the sensory cues only**. Debora explained: *"The first ones we were actively looking at, we were expecting it. I think the last one or two, how many we had, four I think. Yes, probably the first two we were actively looking for it and the last two we saw. I forgot about it and realised the room was cold".* Proactive and reactive approaches were also simultaneously adopted by different members of the same household, signalling a division of tasks in accordance with household's dynamics. Sam and Mara were an example of this, with Sam reactively noticing some attacks due to being more at home and Mara routinely looking for them on the app after arriving late from work.

6.2.2 Experiencing the helper tool and applying remediating actions

Following their initial diagnostic interaction with Squid, as reported in 6.2.1, participants performed the **helper tool checks each time** there was a suspected attack, which was an artefact of the study design, with most completing this as soon as the attack was suspected. As they gained more experience mitigating the first attacks, however, many participants considered the use of the tool a way to **verify what they perceived they could already do, providing the benefit of reassurance**. For example, Kevin explained: "*I used the tool every time, even I know how I should do it, but I would still use it to make sure. I just make everything right."*

In the context of their growing confidence, and other demands in the home (e.g., children's bedtime), several participants began to interact with the tool's **question-checks quickly**. The middle column of Figure 6 sheds light into whether participants answered the tool's checks accurately and, thus, if they received the correct attack type assessment and action. A1 was identified correctly 90% of the time. Despite a few participants answering the question-checks correctly during A2 and A3, overall, they were incorrect 60% of the time and were most often led toward the A1 pathway. Given these findings, it is not surprising, that one key issue reported was the opinion that the **question checks, and recommended steps were the same each time**. For instance, Maya explained: *"I was using the helper tool but*

then it says the same thing, always, it's the same information all the time so I learned the information, at some point, and, after a while, I checked, myself, these pages". Although interestingly, three participants who reported this did manage to identify one A2 attack correctly, but did not demonstrate an awareness of observing this different pathway. This perception reinforced a consensus amongst those who used Squid the most that the helper tool should have been less prominent with the passage of time. Only one participant, Carrie, recognised the possibility of different attack types and thus the importance of using the helper tool to disambiguate them: "*my only hesitation of doing that was because you said (during the training) there were different routes through the helper tool and there might be different things happening so that's why I thought I needed to use the helper tool to make sure I was getting to the right outcome, if you like.*"

Additionally, during one third of the attacks and in all ten households, participants **reset the affected profile** (i.e., completed the remediation action) **before using the helper tool** to diagnose the attack type (indicated with *****), as Figure 7 illustrates. Many, like Debora, were driven to take the action due to their sense of efficacy and their confidence in having acquired transferable skills: "*We used the help in the beginning in the first time. afterwards we didn't need them anymore. It was quite straight forward and that's why we kept doing the same."* Additionally, there were **specific and sometimes persistent situational reasons** for performing the time-efficient action of resetting before using the helper tool, which required more time, e.g., being in the middle of making dinner, having a work deadline, or feeling unwell. Simone explained: "*Then there were times when I would diagnose things on days [reset the profile] I was working and so I was busier and I thought, "Right, I'll remember, I'll do that later." Then I got the flu and then, kind of, every… I saw things and then I thought, "I'll do it another time." It felt like it'd [using the helper tool] get later and later every time.*"

Finally, after users had completed the question-checks, the helper tool presented a **video** connecting the visual indicators with an attack type and recommending a remedial action. For A1 & 2 this consisted of resetting the profile whereas A3 required a switch from the auto to manual mode. Overwhelmingly, participants reported watching the video only once, after the first attack, which was in all but one case for A1. This reinforced participants' perception that resetting the profile attack in their first experience was the appropriate action for all attacks. Only one participant, Carrie, watched the videos throughout the study duration and explicitly recognised their benefits: *"[The video] was almost the most useful part, actually. Yes, because it just talked you through everything so I thought that was the most useful bit of it".*

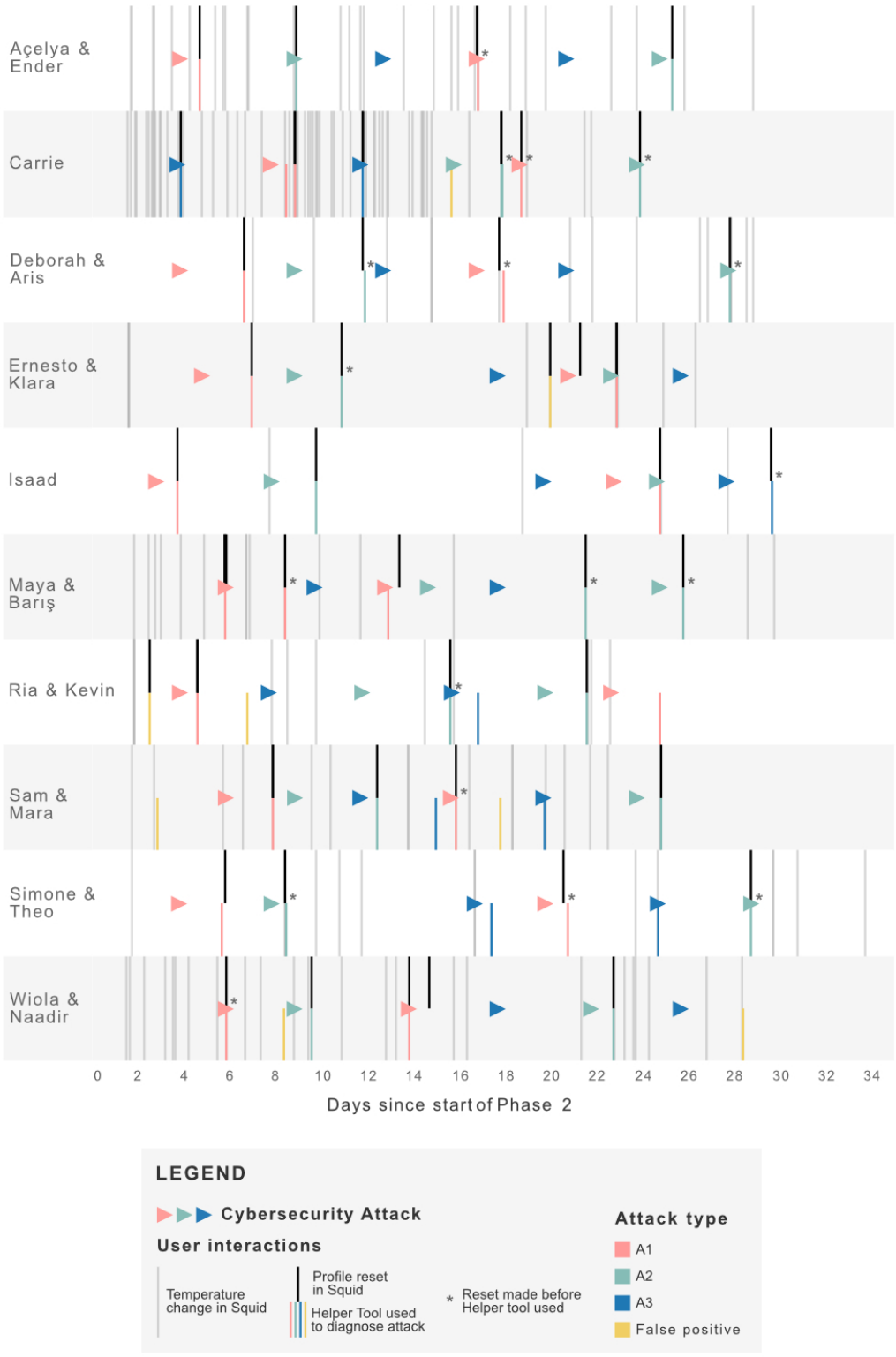


Figure 7. Timeline of attacks and user interactions with Squid and the helper tool

6.3 RQ3: How do people negotiate cybersecurity in the context of AI?

6.3.1 Learning one's normal in the context of AI

All participants reported some level of understanding of what a "normal" temperature was for their Squid, which they checked using its features. Three participants elaborated on how they

expected the AI to **set the temperature within a range** that was acceptable to them based on their previous temperature inputs. Temperatures that contradicted this expectation were perceived to be an indicator of an attack. For example, Mara noticed a drop in temperature, but she did not deem it "drastic" enough to be an attack. In making sense of temperature as a range, Carrie speculated how the AI finds an average between all her temperature inputs: "*I got confused that week because I was looking at the average temperatures and how I'd been setting them and realised that I'd been… I got confused because it was setting it quite low, and I wasn't sure whether that was the AI having learned from me or whether it was actually an attack. It wasn't an attack. It was the AI*". As this quote also illustrates, the fluctuating temperatures set by the **AI introduced constant ambiguity** and required participants to assess if a target temperature was normal or anomalous, which was possible if participants continued to maintain their engagement with Squid: "*Yes, because you see how these things, how they should be so for example how it should look. When something is off, you understand immediately that there something happened [Debora]".*

Participants' intuitive knowledge about their AI temperature ranges, however, was disrupted during the study. In contrast to the efforts they had taken to calibrate Squid's heating profiles in the first phase, as Figure 7 shows, eight households did not consistently undertake retraining after resetting a profile that might have been compromised. A range of reasons were reported during the interviews, such as not realising that re-training was necessary for optimal use or opting for the baseline model where it was perceived to operate within an acceptable level considering the effort involved to retrain a heating profile. For instance, Carrie regularly monitored Squid via the notification logs and felt that the baseline model would set the temperature within an acceptable range for her: "*Yes, I stopped trying to teach it quite so much because I realised, actually, it was setting it close to… The result was closer to what I wanted than 22 degrees in any case*".

Participants' generally **weak understanding of the baseline model** – and its more extreme temperature settings – made it unclear to them whether a temperature drop was the outcome of the AI heating model or a genuine attack, with many struggling to disambiguate this. Maya reported resolving this challenge by inspecting the slope steepness of the AI summary to confirm the normality of AI. Others, however, remained confused or questioned their abilities, such as Carrie who used the helper tool to verify a false positive, or Simone and Theo who were unable to explain why an infected profile (that had not been retrained) was not behaving as they expected: "*I still didn't quite understand it because I thought we'd, sort of, calibrated it in the first few weeks but it kept on- Even though we kept on setting it at 18°, it kept on bringing it down and down... I didn't judge it as being an attack because it was always trying to set it at like 15°".*

The importance of understanding "one's normal" in AI also extended to the relationship between price and temperature. Specifically, the gauge summarised this relationship and acted as a visual indicator of an attack (with red sections communicating to a user if their price sensitivity was abnormally high, or low). Guided by the helper tool, many participants used the gauge as a shortcut to initially glean if there was an attack. Nonetheless, while using Squid three participants activated the gauge's red sections as part of their normal heating practices. Two of them, whose gauge was sometimes pointing to red, used this as a primary method to detect attacks showing little recognition of how the gauge **reflected what constituted as normal** for their household. Similarly, others only considered the baseline model adequate due to its target temperature, ignoring or misinterpreting the unusual price sensitivity indicated by the gauge. In relying on this indicator, it was thus critical to have some prior understanding of how it reacted in response to one's everyday living, which Carrie expressed: *"The gauge was and wasn't useful because I think, also, I was setting the temperature. It was saying I was very sensitive to price changes so it was quite often across*

in the red and that's partly… I guess, if you were using the system in reality, you'd set it to hit the temperature you would want and I thought… I think that's where I found things a bit difficult to spot because it was often in the red".

6.3.2 Lack of prior knowledge to inform price benchmarks

The price dimension of the AI model also raised challenges as participants entered the research without any prior **price benchmarks**, in support of decision making. This gap was negotiated in several ways. Lacking a reference point, twelve participants reported looking for irregularities in the temperatures, discarding price altogether. Theo captured this idea arguing that "*I just found it too abstract about what represented high and low for a given time of day."* Two participants, Sam and Kevin, who were not responsible for their heating bills and thus had little engagement with energy prices, shared extreme or inaccurate assumptions of benchmarks for assessing price irregularity: "*I know pretty much the prices that it should be, it shouldn't be £1 or £100 or whatever*" *[Sam]*. Recognising her limited knowledge of prices, and informed by the initial training, Simone employed the side panel to develop an ongoing understanding of the price ranges.

With most noting a lack of understanding in price benchmarks and their ranges, A3 was the least detected attack, which Ender explained "*I don't know the numbers. I mean, 20, it can be okay, but 50, maybe it can be, I don't know. I mean, I don't have a reference point, so probably I would never […] detect that (A3)*". Moreover, it is noteworthy that three participants reported noticing the low temperatures, reaching the conclusion that the high prices they observed offered a plausible explanation for the attack: "*So I checked the overall temperature/price ratio and I thought,'maybe it's thinking the price is really high so it had to make it like 7C°', so I was thinking it's the logic of the AI, basically*" *[Maya]*. Given the plausibility of the high prices observed in A3, even those who triggered the helper tool upon observing the attack, such as Simone, reported experiencing doubt. The interviews also highlighted a second challenge relating to the **expectedness** of A3. Several participants expressed surprise when hearing or corroborating the price attack. As two of them shared, this was due to the regulation they expected on energy pricing, and thus their reliance on the provider in how they were billed, which extended to Squid.

6.3.3 Actions for resolving uncertainty and anxiety

During the study, five households reported eight false positives against 48 correct diagnoses. The false positives were partly caused by the ambiguity introduced when using the baseline AI models as discussed under 6.3.1. However, there were a range of other reasons leading participants to suspect an attack. In several households there were standard hardware issues that occasionally triggered more buzzing in the smart valve, creating confusion since buzzing was an auditory cue of a cyber-attack. Another household felt the room was colder than normal, which was caused by the physical distance between the Squid's valve temperature reading (that was attached to the heat source) and the central thermostat that was in a different room requiring the thermostat to be set higher than expected to align with one's sensory preferences. All of those experiencing a false positive successfully used the **helper tool** to establish that these events were not cyber-attacks.

Alongside the helper tool's value in these occasions, the **action of resetting** the infected profile emerged as particularly important. As reported in 6.3.2, all the households used the reset button following their suspicion of an attack. This provided some respite before they could engage with the procedures of the helper tool. According to Wendy and Naadir, the ability to act alleviated their anxiety, with Ernesto discussing the profile reset as a "low-cost action." Additionally, a few participants used this action to resolve uncertainty. The reset button was vital when participants were unsure if there was an unfolding attack, such as in

the case of Ender whose gauge had a fault during the A2 attack using the button to restore the system back to its original state and thus verify the attack. After receiving a false positive outcome from the helper tool, the reset button was also used by Carrie to ease her own anxiety: "*I think I'd got into my head that you needed to reset to be safe and sure, put everything back and then you can start again.*"

7 Discussion

This research set out to explore a new cybersecurity intervention designed to support users to recognise, understand, and address integrity cyber-attacks on the AI component of a smart device, using an AI-enabled heating system called Squid as an exemplar to inform the intersection of AI domestic technologies and cybersecurity. Below we reflect on the design implications suggested by our findings. We draw from lessons learned in the context of Squid to inform other AI-enabled devices used in the home which could be vulnerable to similar attacks. We are careful to situate our findings within the context and limitations of this research. An important point of consideration is that the proposed scenario did not involve sensitive personal data, making the consequences of an attack potentially less concerning than other AI-enabled devices such as voice assistants or cameras. Additionally, participants were primed to proactively look for attacks, whereas real-world users are unlikely to be regularly checking for cyber-attacks. Therefore, in this section we consider what we can learn given our research design.

7.1 Multimodal indicators are effective at raising attention to cyber-attacks

Previous research has highlighted that AI-enabled device malfunctions and unexpected technology behaviours (Huijts et al., 2023; Rostami et al., 2022) introduce inherent ambiguities and consequently pose challenges for users, who justifiably struggle to differentiate them from cyber-attacks. It is thus crucial to design perceivable indicators that can make users aware of a potential cyber-attack (Kuzlu et al., 2021), but what these may be for AI attacks and how salient they should be for users to detect them remains an open question.

Our study advances this research gap by demonstrating that the *visual* indicators (some reliant on Explainable AI techniques e.g., the gauge, the notification log), were particularly effective and valuable to the participants, supported by the fact that A1 and A2 were identified most of the time. However, it was also observed that these were rarely perceived in isolation. Instead, participants relied on diverse combinations of indicators – a finding we return to in 7.3. Whereas designing multiple visual indicators may be one approach to mitigate cyber-attacks, it is also important to recognise that oftentimes visual indicators can be explicitly crafted to hide in plain sight as part of the attack (Comiter, 2019; Kuzlu et al., 2021), which suggests the importance of considering the introduction of other modalities.

Speaking to this question, it is instructive to consider the effectiveness of the sensory and auditory indicators activated during the cyber-attacks simulated in the study, and the trade-offs we identified. The *sensory* indicator (i.e., feeling cold) enhanced the cyber-attack detection in A1 and A2, but its effectiveness depended on the persistence of the attack and its ongoing impact on the household environment. Participants tended to perceive a colder environment if the attack was long-lasting and/or if someone was present in the affected room or its vicinities. Therefore, properties of the physical environment and patterns of room occupancy limited the extent of the felt experience for some, leading to households having a varied experience with this indicator in accordance with their routines and living

circumstances. The *auditory* indicator proved to be unreliable in most of the households for having a shorter-term effect and requiring greater proximity to the sound source. Additionally, in some of the households the auditory indicator did not work given the fit of the smart valve with the radiator. Therefore, time, presence and mechanical fit proved to be key factors in the perception or disregard of indicators.

While visual indicators could be perceived remotely and, in some cases, retrospectively, the others had short-term action and relied on physical presence. Nevertheless, *sensory,* and *auditory* indicators were still relevant in combination with visual indicators and are likely to be more prominent in a real-life situation in which a smart heating system would be installed in the entire house instead of a single radiator. Auditory indicators could be a low-cost addition to a set of cybersecurity indicators, but it is relevant to note that device manufacturers would have likely ensured the valve was as quiet as possible so as not to disturb the user, without recognising the value of this modality. This highlights a trade-off that needs to be made with how auditory indicators are designed, i.e., between enabling the device to be seamlessly embedded into the home and alerting the user of what it is doing.

Additionally, while visual indicators proved to be the most promising in this study, these were effective whilst participants were proactively looking for attacks – a behaviour induced by the design of the study and participants' expectations of upcoming attacks. Contrasting to this, prior research has shown that users tend to disengage with AI-enabled devices after the initial set up period (Vasalou et al, 2024; Jensen et al., 2017). A reactive approach would be more representative of how people use technology in everyday life and in this case the most reliable indicator would be the sensory one, i.e. feeling the cold, which is a negative consequence of the attack. Owing to the limitations of the auditory and sensory indicators noted above, future work could explore how to **make visual indicators more prominent** to support reactive responses, such as embedding them into the physical device. For example, the physical smart valve screen could flicker or glow to indicate an anomaly, prompting the user to visit the associated web app for further investigation.

It is possible that the design and function of an AI-enabled device, as well as its location in the home, may impact on the perception of different indicators, i.e., each device and its context of use may draw attention to different indicators. In recognition of this, our findings highlight the relevance of including a diverse set of indicators in AI-enabled devices (visual, auditory, sensorial) that can make an attack perceivable amidst the noise of everyday life and cater for users' differential environment, routines, perceptions, and abilities. Given the consistent spotting of attacks, particularly those with a pervasive impact on the environment, our study supports the paradigm of *humans as sensors* in the detection process (Heartfield and Loukas, 2018b). There is potential for future work to explore how to make the link with device manufacturers after these perceptions have been activated.

7.2 Normative behaviour of AI is foundational to make sense of cyber-attacks

Focusing on the *visual* indicators designed within Squid, several were inspired by Explainable AI and intended to communicate abnormalities in the AI model parameters (i.e., temperature and price) to support users' diagnoses of cyber-attacks. This section considers the specific challenges participants encountered when engaging with these features, aiming to contribute to the nascent research at the intersection of human-centred cybersecurity and interaction design of Explainable AI.

Our study raises a first challenge related to the AI's parameters and their relevance within people's everyday lives. In our research, this is related to the disclosure of simulated energy prices. Given households' overall priority to sustain their thermal comfort (reported in Vasalou et al., 2024), and the lack of tangible impact of the prices used by the study's design,

participants engaged with price in a cursory way and discarded this parameter. Consequently, many failed to notice the A3 attack on the pricing model, and those who identified the heightened prices lacked the benchmark to appraise this as abnormal. Whilst the weak relevance of price was specific to the controlled nature of this study, our findings highlight the importance of ensuring Explainable AI is based on models that incorporate parameters holding value and meaning to users, an issue that has broader relevance to AI-enabled technology design beyond smart home heating.

A second challenge users encountered stemmed from their lack of a mental model of what constitutes "normal" AI behaviour in the context of their use of Squid, which is a consideration applicable to any AI-enabled devices that may invite user engagement with their AI. While this applied to both factors of temperature and price alike, we focus our discussion on temperature given participants' discard for price. When heating their homes prior to Squid's introduction, participants typically controlled their heating by selecting a specific target temperature aligned with their contextual preferences. In contrast, the AI introduced an understanding of *target temperature as a range,* as opposed to a *precise value*. Thus, in order to identify a cyber-attack, it was necessary to acquire an understanding of what temperature ranges constituted as normal. To achieve this, it was vital not only to inspect the AI, but to also relate it to one's previous inputs. With many users not engaging in this cycle of input-inspection, our study shows there was a heightened risk of incorrectly concluding there was a cyber-attack. A future implication of triggering a higher frequency of false positives could be that some users unnecessarily experience anxiety, while others could become desensitized over time and overlook genuine cyber-attacks. Furthermore, our study shows that failing to understand what constitutes normal AI behaviour can contribute to users misinterpreting summative Explainable AI visualisations. In our research this was exemplified by the gauge and its pre-determined price sensitivity ranges that visually alerted users of extreme sensitivities associated with a cyber-attack, which may have been nonetheless within the normal range for the household's heating practices.

Thus, despite the study using the simplest case of an AI model (involving a glass box algorithm) users faced interpretive challenges when interacting with Explainable AI, which was shown to lead to both over/underestimations of cyber-attacks. What is clear from our findings is that **an ongoing understanding of AI model parameters is foundational for users to diagnose a cyber-attack**. Moreover, the broader finding that false positives can be triggered by different causes, including users' lack of understanding of the AI models, underscores the importance of design that can enhance smart device transparency and minimise the causes of false positives.

7.3 Designing to support diagnosis with assisted and transferable learning

Understanding the potential cause of a cyber-attack and possessing the skill to respond to it have been recognised as key to enabling the user to enact an effective cybersecurity role (Frik et al., 2019). This is particularly pertinent as cyber-attacks might have different causes that require different actions to be resolved (Pitropakis et al., 2019). As such, one of the aims of the cybersecurity intervention reported in this research was to bolster users' understanding and skill acquisition. This was achieved through guidance from a bespoke helper tool designed to support users to correctly interpret patterns of indicators (inclusive of Explainable AI), exposing different attack causes whilst providing recommendations for differentiated action.

In contrast to the helper tool's design intentions, participants perceived multiple indicators related to the attacks but were not necessarily aware of their *variety*. Instead of acquiring the skill to use the helper tool to disambiguate the type of cyber-attack and course

of action, each participant developed their own routine. This involved the examination of a fixed set of indicators across all three types of cyber-attacks and the application of the same action (to reset a Squid profile), which had been practiced as part of the initial training session with A1 and was also applicable to A2. Further examination of how the helper tool question-checks were used (see Figure 6) showed that compared to A1 where diagnosis was 90% accurate, for A2 and A3 this dropped to 45% and 10% respectively, with most participants incorrectly diagnosing the remaining two attacks as A1.

These forms of engagement and practices related to the cybersecurity intervention introduced a vicious cycle: *participants followed the incorrect pathway within the helper tool without being aware when they misdiagnosed the cyber-attacks for A1. This strengthened the perception that cyber-attacks and their remedial actions were always the same, and reinforced the idea that the best strategy was to rely on a fixed set of indicators to diagnose all attack types*. The consequences of this cycle were most vividly exemplified in A3. It introduced Squid users to the idea that attacks could target different components of the smart device ecosystem, and occur at the level of the device manufacturer, notably a frequent type of attack encountered in the consumer context. From the four households who registered an A3, none had taken the correct remedial action and when discussing these results retrospectively with these households they expressed surprise at what had caused it.

Drawing from our behavioural data, these engagement patterns could be attributed to participants generalising what they learned in their training with A1 to the other cyber-attacks, aligned with previous research indicating that people tend to reapply known cybersecurity strategies to new devices (Bouwmeester et al., 2021; Zeng et al., 2017). Accordingly, participants generalised the same strategy across the three attacks developing a singular, inaccurate mental model of cyber-attacks. In contrast to attacks which do not share the same cause and where transferability of learning could expose users to risk, we suggest that **attacks of the same type (as were A1 and A2) that share similar remedial actions, causes and indicators, could provide a promising avenue to develop training that elicits transferable learning**. Another explanation, supported by the study's qualitative findings, was that the helper tool's final video summary, which contained crucial information about the causes of each attack diagnosed and its suitable remedial action, was ineffective. Drawing from the ineffectiveness of the video summary, future helper tools could **build up the causal explanation of the attack whilst the user answers the diagnostic questions**.

In addition to how people cognitively engaged with the helper tool's design and the design implications raised, the home context had a profound impact on participants' general level of engagement with most citing demands of everyday home life as key reasons to this. This weak engagement reflects what has previously been observed by other researchers (Zimmermann and Renaud, 2019). We contend that user assistance in the form of a helper, is vital to protect users particularly against novel attacks, yet as our study shows users' time to deeply engage can be limited and contingent to what is happening in their home at the time. Moreover, the need to monitor and understand cyber-attacks on AI-enabled devices runs counter to the expectation many have that automation saves time. To address this challenge, designers of these devices could consider how to incorporate helper tools that guide users to **troubleshoot their devices and investigate attacks in semi-automated ways to reduce effort**, e.g., checking through pages of notification logs, making potential abnormalities more salient. Such interaction design strategies could reduce the time-consuming analysis our intervention required without undermining user understanding.

Considering the helper tool as a form of 'training' to recognise attacks, one might anticipate users accessing it less as they grow more confident in their own abilities. Our study evidences the range of reasons that can disrupt this training, raises implications at the level of cybersecurity intervention design, and reveals the importance of carefully considering

people's perspectives on how these procedures integrate into their daily routines. As shown in 7.1 the helper tool successfully drew attention to a range of indicators that might have otherwise been overlooked, despite failing to convey the idea that AI cyber-attacks can take different forms that require differentiated actions. To address these caveats, future research could actively engage users in co-designing such interventions whilst grounding them in theories of how people learn.

7.4 Impactful actions for AI cyber-attacks

A key challenge reported in previous literature relates to the complicated, or onerous, procedures required to resolve cyber-attacks that constrain user agency (Bouwmeester et al., 2021; Rodriguez et al., 2022). Contrasting with this, the main remedial action in two of the three cyber-attacks within Squid was to reset the profile, restoring the AI model(s) through a single click. Prior research reports that users prioritise addressing the immediate consequences of a cyber-attack over engaging and dealing with the underlying cause(s) (Frik et al., 2019). In line with this, our findings showed that in a third of the cyber-attacks participants undertook the remedial action before engaging with the helper tool, which would have allowed them to assess whether the action was appropriate. This pre-emptive behaviour was likely accentuated by the frequency of cyber-attacks in the study. The impetus to take action was driven by a need to resolve uncertainty when it was not clear whether a cyber-attack was ongoing, and to alleviate the anxiety this triggered. Since home routines competed with people's ability to engage with the helper tool, participants also believed the resetting of profiles to be a low-effort action due to it being quick to perform and restore the system.

Contrasting with this perception, however, resetting the profile came with the imperative of time and effort. It raised the need for participants to retrain the AI model in line with their household's heating preferences, at the risk of introducing ambiguity that made it subsequently more difficult for new attacks to be identified (see 7.2). Whilst simple restorative actions like these (i.e., reset) can offer immediate comfort, they can also be redundant introducing hidden effort on the part of the user. In the context of Squid, a more suitable immediate action for managing ambiguity would have been to put the device in manual mode (thus removing control from the AI), interrupting the potential cyber-attack and affording time to act. Therefore, particularly in scenarios where the causes of cyber-attacks may be challenging to identify, our findings emphasise a consideration for **designing "universal actions"** that offer users with different technical skills and availability the capability to pause or restore the impacts of cyber-attacks, alongside the need to **educate them about the costs/consequences of the different actions** available to mitigate cyber-attacks.

7.5 Limitations and future work

Conducting cybersecurity research in the wild is challenging given the low incidence of cyber-attacks in everyday life. In this work we addressed this challenge by emulating attacks that could be subsequently mitigated through purposefully designed features embedded in an AI-enabled device. Given this methodological approach, inevitably participants experienced a sense of vigilance and tuned into the cyber-attacks. As described under 6.2.1, whilst in the first week most households were looking out for attacks, some reported a decline in attending to the cyber-attacks over the course of the study. Future work that draws on a similar research design may seek to extend the study timespan to foster this habituation and enhance naturalistic behaviours. Furthermore, to investigate the impact of the multimodal indicators, we took the decision to time the attacks to when participants reported being typically at

home. Despite these efforts, as Figure 7 suggests, there was often a time lag between an attack and participants' response, owing to participants' not being at home, or not present in the room at the time of the attack. Therefore, considering the practical challenges involved when researching cyber-attacks in situ, we suggest that timing attacks to participants' self-reported presence in the home is likely to give access to both synchronous and asynchronous experiences of cyber-attacks.

8 Conclusion

With the proliferation of AI-enabled devices in the home, cyber-attacks targeting their AI component are poised to become increasingly common, yet they have been neglected in the literature to date. At the same time, manufacturers are generally reluctant to outwardly acknowledge their devices are vulnerable to attacks (Turner et al., 2022) highlighting the importance to follow a responsible design approach (alongside the need for strong regulation). Our research posits that users of AI-enabled devices can play an important role in protecting themselves against these attacks, but to do so, they need to be supported through intentionally designed devices and appropriate cybersecurity support tools. Moving away from seeing the user as the problem to recognising the user as part of the solution (Roba Abbas et al., 2023), our paper developed a novel cybersecurity intervention to support users to identify, diagnose and mitigate integrity cyber-attacks on the AI component of their smart devices. Employing a case study of a bespoke smart heating device, Squid, we conducted fieldwork with ten households. This allowed us to evaluate the intervention over a course of four weeks during which each household experienced six cyber-attacks over three types.

Our findings advance the imperative to support users to keep cybersecure during attacks on the AI through the following contributions. *First*, we show that multimodal indicators designed across the AI-enabled device and its ecosystem are a reliable technique to raise people's awareness of cyber-attacks. One implication our study introduces for AI-enabled devices is the need for visual indicators embedded on the physical device to alert the user to visit the information-rich environment of a smart device app. *Second*, we find that shifting the user from a sensing role to one that involves actively troubleshooting the type, cause, and action pertinent to a given cyber-attack must avoid cognitive overload and fit in the routine of home life. Crucially, our research indicates that without these foundations users may be exposed to more false positives and/or reach inaccurate cyber-attack diagnoses, raising the likelihood of a range of harms, from increased anxiety to desensitisation of genuine cyber-attacks. In support of this direction, we underscore the following design implications: embedding causal explanations of the attack into the troubleshooting process, highlighting what to look for within text-based Explainable AI (e.g. logs), and where possible supporting transferrable learning across cyber-attacks that share causes/mitigative actions akin to traditional cyber-hygiene preventive approaches (e.g. not re-using passwords). *Third*, and relatedly, we advance the importance of the AI-enabled device offering universal mitigation actions that can support the user's efficacy and raise transparency with regards to the costs introduced by the action. One avenue we highlight is the possibility of design features that can pause the AI affording users with time to reflect. These findings are contextualised to the distinctive challenges raised by AI, in particular, where the diagnosis of cyber-attacks depends on how well users understand a fluid range of outputs in relation to their own behaviour, and the direct experience they have with the AI's parameters, a consideration that remains a challenge depending on the algorithmic approach taken.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council through the CHAI Project [EP/T026812/1]. We gratefully acknowledge the ten households who participated in this research and also Rytis Vensolovas for is contribution to Squid's interaction/visual design.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M., 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '18: CHI Conference on Human Factors in Computing Systems, ACM, Montreal QC Canada, pp. 1–18. https://doi.org/10/gfzzgc

Alan, A.T., Shann, M., Costanza, E., Ramchurn, S.D., Seuken, S., 2016. It is too Hot: An In-Situ Study of Three Designs for Heating, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16. Association for Computing Machinery, New York, NY, USA, pp. 5262–5273. https://doi.org/10.1145/2858036.2858222

Benton, L., Vasalou, A., Turner, S., 2023. Location, Location, Security? Exploring Location-Based Smart Device Security Concerns and Mitigations within Low-Rent Homes, in: Proceedings of the 2023 ACM Designing Interactive Systems Conference. Presented at the DIS'23, pp. 1060–1077.

Bouwmeester, B., Rodríguez, E., Gañán, C., van Eeten, M., Parkin, S., 2021. "The Thing Doesn't Have a Name'': Learning from Emergent Real-World Interventions in Smart Home Security, in: Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021). pp. 493–512.

Chalhoub, G., Kraemer, M.J., Nthala, N., Flechais, I., 2021. "It did not give me an option to decline": A Longitudinal Analysis of the User Experience of Security and Privacy in Smart Home Products, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '21: CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, pp. 1–16. https://doi.org/10.1145/3411764.3445691

Chen, Q., Romanowich, P., Castillo, J., Roy, K.C., Chavez, G., Xu, S., 2021. ExHPD: Exploiting Human, Physical, and Driving Behaviors to Detect Vehicle Cyber Attacks. IEEE Internet Things J. 8, 14355–14371. https://doi.org/10.1109/JIOT.2021.3069951

Comiter, M., 2019. Attacking Artificial Intelligence (No. 8), Belfer Center Paper.

Committee, C., Media and Sport, 2022. Connect tech: smart or sinister (No. Tenth Report of Session). House of Commons.

Dourish, P., Anderson, K., 2006. Collective Information Practice: Exploring Privacy and Security as Social and Cultural Phenomena. Human Computer Interaction 21, 319–342. https://doi.org/10.1207/s15327051hci2103_2

Ehrenberg, N., Keinonen, T., 2021. Co-Living as a Rental Home Experience: Smart Home Technologies and Autonomy. Interaction Design and Architecture 50, 82–101. https://doi.org/10.55612/s-5002-050-005

Frik, A., Nurgalieva, L., Bernd, J., Lee, J.S., Schaub, F., Egelman, S., 2019. Privacy and Security Threat Models and Mitigation Strategies of Older Adults, in: Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019). pp. 21–40.

Guardian, 2023. Energy bills: 13m British homes 'did not turn on heating when cold last winter.'

Hammi, B., Zeadally, S., Khatoun, R., Nebhen, J., 2022. Survey on smart homes: Vulnerabilities, risks, and countermeasures. Computers & Security 117, 102677. https://doi.org/10.1016/j.cose.2022.102677

Heartfield, R., Loukas, G., 2018a. Detecting semantic social engineering attacks with the weakest link: Implementation and empirical evaluation of a human-as-a-security-sensor framework. Computers & Security 76, 101–127.
Heartfield, R., Loukas, G., 2018b. Detecting semantic social engineering attacks with the weakest link: Implementation and empirical evaluation of a human-as-a-security-sensor framework. Computers & Security 76, 101–127. https://doi.org/10/gdv3d5
Heartfield, R., Loukas, G., Budmir, S., Bezemskij, A., Fontaine, J.R.J., Filippoupolitis, A., Roesch, E., 2018. A taxonomy of cyber-physical threats and impact in the smart home. Computers & Security 78, 398–428.
Huijts, N.M.A., Haans, A., Budimir, S., Fontaine, J.R.J., Loukas, G., Bezemskij, A., Oostveen, A., Filippoupolitis, A., Ras, I., IJsselsteijn, W.A., Roesch, E.B., 2023. User experiences with simulated cyber-physical attacks on smart home IoT. Personal and Ubiquitous Computing.
Jensen, R.H., Kjeldskov, J., Skov, M.B., 2018a. Assisted Shifting of Electricity Use: A Long-Term Study of Managing Residential Heating. ACM Transactions Computer Human Interaction. 25, 1–33. https://doi.org/10.1145/3210310
Jensen, R.H., Strengers, Y., Kjeldskov, J., Nicholls, L., Skov, M.B., 2018b. Designing the Desirable Smart Home: A Study of Household Experiences and Energy Consumption Impacts, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '18: CHI Conference on Human Factors in Computing Systems, ACM, Montreal QC Canada, pp. 1–14. https://doi.org/10.1145/3173574.3173578
Jeong, J.J., Oliver, G., Kang, E., Creese, S., Thomas, P., 2021. The current state of research on people, culture and cybersecurity. Personal and Ubiquitous Computing 25, 809–812.
Kuzlu, M., Fair, C., Guler, O., 2021. Role of Artificial Intelligence in the Internet of Things (IoT) cybersecurity. Discov Internet Things 1, 7. https://doi.org/10.1007/s43926-020-00001-4
Loukas, G., 2015. Cyber-physical attacks: A growing invisible threat. Butterworth-Heinemann.
Meneghello, F., Calore, M., Zucchetto, D., Polese, M., Zanella, A., 2019. IoT: Internet of Threats? A Survey of Practical Security Vulnerabilities in Real IoT Devices. IEEE Internet Things J. 6, 8182–8201. https://doi.org/10.1109/JIOT.2019.2935189
Mennicken, S., Vermeulen, J., Huang, E.M., 2014. From today's augmented houses to tomorrow's smart homes: new directions for home automation research, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Presented at the UbiComp '14: The 2014 ACM Conference on Ubiquitous Computing, ACM, Seattle Washington, pp. 105–115. https://doi.org/10.1145/2632048.2636076
Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), pp.22071-22080.
Nagaraju, V., Fiondella, L., Wandji, T., 2017. A survey of fault and attack tree modeling and analysis for cyber risk management, in: 2017 IEEE International Symposium on Technologies for Homeland Security (Hst). IEEE, pp. 1–6.
Nicholls, L., Strengers, Y., Sadowski, J., 2020. Social impacts and control in the smart home. Nature Energy 5, 180–182.
Peters, U., 2023. Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. AI and Ethics 3, 963–974.
Pink, S., Leder Mackley, K., Morosanu, R., Mitchell, V., Bhamra, T., 2017. Home: ethnography and design, Home. Bloomsbury Academic, London ; New York.
Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., Loukas, G., 2019. A taxonomy and survey of attacks against machine learning. Computer Science Review 34, 100199. https://doi.org/10.1016/j.cosrev.2019.100199

Rahman, T., Rohan, R., Pal, D., Kanthamanon, P., 2021. Human Factors in Cybersecurity: A Scoping Review, in: The 12th International Conference on Advances in Information Technology. Presented at the IAIT2021: The 12th International Conference on Advances in Information Technology, ACM, Bangkok Thailand, pp. 1–11. https://doi.org/10.1145/3468784.3468789

Roba Abbas, K.M., Pitt, J., Vogel, K.M., Zaferirakopoulos, M., 2023. Artificial Intelligence (AI) in Cybersecurity: A Socio-Technical Research Roadmap. The Alan Turing Insitute.

Rodriguez, E., Fukkink, M., Parkin, S., van Eeten, M., Ganan, C., 2022. Difficult for Thee, But Not for Me: Measuring the Difficulty and User Experience of Remediating Persistent IoT Malware, in: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P). Presented at the 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), IEEE, Genoa, Italy, pp. 392–409. https://doi.org/10.1109/EuroSP53844.2022.00032

Rostami, A., Vigren, M., Raza, S., Brown, B., 2022. Being Hacked: Understanding Victims' Experiences of IoT Hacking, in: Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022). pp. 613–631.

Slupska, J., Dawson Duckworth, S.D., Ma, L., Neff, G., 2021. Participatory Threat Modelling: Exploring Paths to Reconfigure Cybersecurity, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. Presented at the CHI '21: CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, pp. 1–6. https://doi.org/10.1145/3411763.3451731

Spero, E., Biddle, R., 2019. Security Begins at Home: Everyday Security Behaviour and Lessons for Cybersecurity Research, in: Proceedings of the 26th Conference on Pattern Languages of Programs. pp. 1–9.

Still, J.D., 2016. Cybersecurity needs you! interactions 23, 54–58. https://doi.org/10.1145/2899383

Turner, S., Nurse, J.R.C., Li, S., 2021. When Googling it doesn't work: The challenge of finding security advice for smart home devices. pp. 115–126. https://doi.org/10.1007/978-3-030-81111-2_10

Turner, S., Pattnaik, N., Nurse, J.R.C., Li, S., 2022. "You Just Assume It Is In There, I Guess": Understanding UK Families' Application and Knowledge of Smart Home Cyber Security. Proceedings of ACM Human Computer Interaction. 6, 1–34. https://doi.org/10.1145/3555159

Vasalou, A., Gauthier, A., Serta, A., Besevli, C., Turner, S., Payler, R., Gill, R., McAreavey, K., Benton, L., Loukas, G., Liu, W., Beneito-Montagut, R., 2024. In Pursuit of Comfort: An Exploration of Smart Heating in Everyday Life. International Journal of Human-Computer Studies.

Vuong, T.P., Loukas, G., Gan, D., Bezemskij, A., 2015. Decision tree-based detection of denial of service and command injection attacks on robotic vehicles, in: 2015 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1–6.

Warford, N., Matthews, T., Yang, K., Akgul, O., Consolvo, S., Kelley, P.G., Malkin, N., Mazurek, M.L., Sleeper, M., Thomas, K., 2022. SoK: A Framework for Unifying At-Risk User Research, in: 2022 IEEE Symposium on Security and Privacy (SP). Presented at the 2022 IEEE Symposium on Security and Privacy (SP), IEEE, San Francisco, CA, USA, pp. 2344–2360. https://doi.org/10.1109/SP46214.2022.9833643

Zeng, E., Mare, S., Roesner, F., 2017. End User Security & Privacy Concerns with Smart Homes, in: Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017). pp. 65–80.

Zimmermann, V., Renaud, K., 2019. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. International Journal of Human-Computer Studies 131, 169–187. https://doi.org/10.1016/j.ijhcs.2019.05.005