# *Simulating human detection of phishing websites:*

## *An investigation into the applicability of ACT-R cognitive behaviour architecture model*

Nick Williams
Department of Computer Science
University of Surrey
Guildford, UK
ngbwilliams@gmail.com

Shujun Li
Department of Computer Science
University of Surrey
Guildford, UK
http://www.hooklee.com/

***Abstract* — The prevalence and effectiveness of phishing attacks, despite the presence of a vast array of technical defences, are due largely to the fact that attackers are ruthlessly targeting what is often referred to as the weakest link in the system – the human. This paper reports the results of an investigation into how end users behave when faced with phishing websites and how this behaviour exposes them to attack. Specifically, the paper presents a proof of concept computer model for simulating human behaviour with respect to phishing website detection based on the ACT-R cognitive architecture, and draws conclusions as to the applicability of this architecture to human behaviour modelling within a phishing detection scenario.**

**Following the development of a high-level conceptual model of the phishing website detection process, the study draws upon ACT-R to model and simulate the cognitive processes involved in judging the validity of a representative webpage based primarily around the characteristics of the HTTPS padlock security indicator. The study concludes that despite the low-level nature of the architecture and its very basic user interface support, ACT-R possesses strong capabilities which map well onto the phishing use case, and that further work to more fully represent the range of human security knowledge and behaviours in an ACT-R model could lead to improved insights into how best to combine technical and human defences to reduce the risk to end users from phishing attacks.**

***Keywords — Phishing, website, security, psychology, human behaviour, cognitive modelling, ACT-R***

## I. INTRODUCTION

### *A. Phishing attack effectiveness – targeting the human*

Phishing – a practice which tricks people into handing over their sensitive data to attackers [1] is an issue which the security industry is a long way from bringing under control. The effectiveness of such attacks is of particular concern, with the SANS Institute for example reporting that 95% of all breaches start with a phishing attack [2]. The key to their success is the fact that they exploit what is often referred to as the weakest link in the system – the human.

Phishing attacks employ a combination of social engineering and "technical subterfuge" [3] to lure unsuspecting users to fake websites controlled by the attacker. These users often then fail to detect the sites as fake, resulting in them placing their trust in the site and entering sensitive information which can lead to significant harm both to individuals and organisations in such forms as financial loss, loss of intellectual property, damage to corporate reputation and identity theft.

### *B. Countering the attack – traditional responses*

The security industry has traditionally sought to counter this threat primarily through the application of technical controls, and yet while technology can play a significant role in defence against such attacks, it cannot solve the problem alone. There are many debates over what the correct blend of 'people, process and technology' controls should be when seeking to counter a security threat [4], and yet it is clear that one cannot ignore the human element in seeking to prevent and detect phishing attacks. Industry's response in this regard has been to deliver user education, training and awareness programs, often following well known standards such as ISO/IEC 27001 [5].

Researchers have been supporting such efforts for many years, with Computer Science and Psychology disciplines in particular contributing to increased understanding of, and solutions to, the problem of phishing attacks in society. But these disciplines, while each individually offering insights and solutions, are also now increasingly joining forces, with leading classic papers such as [6] and [7] strongly arguing for a more holistic approach to security risk management.

### *C. The role of cognitive behaviour analysis*

One specific area in which the two fields have started to work together is the application of human cognitive behaviour research to the field of information security. Leading psychologists [8] have delivered significant advances in our understanding of the human mind, and have developed several mature, well-respected cognitive behaviour frameworks such as ACT-R [9] to model and predict human behaviour in a range of different environments. Within the security space, researchers have started to apply these frameworks to problems such as malware identification [10], but to date these do not appear to have been applied within a phishing context.

This paper reports our investigation into how a computer model of cognitive behaviour might simulate human detection (or acceptance) of phishing websites. In particular, it considers how the ACT-R cognitive behaviour architecture framework and model could be applied to the problem and how, in doing so, we might gain a deeper understanding of how humans behave when faced with a phishing website and the limitations of our current approaches to combatting phishing attacks.

## II. UNDERSTANDING HUMAN DETECTION OF PHISHING WEBSITES – ATTACK VS. DEFENCE

### A. The attacker perspective

To ensure an effective defence against threats to one's assets and systems users are often advised to adopt the security mindset of thinking like the attacker [11]. This requires consideration of the end goal of the phishing attack and the potential strategies, options and constraints which might help or hinder the attacker in seeking to achieve this goal.

*1) The attacker's goal:* In preparing a phishing attack, the attacker is typically seeking to acquire sensitive information, such as passwords, personal data or credit card details. The goal may be fairly general in nature (e.g. "obtain as many credit card details as possible"), or it may be highly specific, targeting one or more individuals or organisations. In defining the goal of a phishing attack, the attacker may also wish to consider other factors, such as desired attack duration, or the level of resources, skills, time and finance available.

*2) Attack strategies:* Having decided upon the goal of the attack, the attacker must then consider how this will be achieved. Phishing attacks have been described as relying on three main strategies: deception, diversion, and exploitation of lack of user knowledge [12]. All of these strategies follow the "carrot" approach in that they seek to convey a positive sense of trustworthiness to the end user. To these we should also consider the "stick" approach, in which techniques such as psychological manipulation are used to pressure the user into entering sensitive details out of fear that they may lose out should they fail to do so.

*3) Phishing website design options:* The attacker must finally decide on how best to implement his or her strategy, leading to consideration of a number of phishing website design options. A wide range of techniques are available, such as: ensuring that the fake website is visually similar to the genuine website or manipulating the browser address bar to create the impression that the site is genuine (deception); use of distracting images and logos to draw the user's attention away from indicators that may reveal the site to be fake (diversion); and domain name cheating, in which the attacker attempts to trick the user into believing that the domain namebelongs to the genuine website rather than the site controlled by the attacker (exploiting lack of user knowledge).

*4) Attacker limitations:* The opportunities for tricking end users into submitting sensitive infromation into a phishing website are many and varied, and there is no simple way of guaranteeing that a given site is genuine or fake. Nevertheless, the attackers themselves are limited to the extent that they cannot develop phishing websites which will fool all end users all of the time. Given the right knowledge and applying the right behaviours, it is possible for end users to greatly increase their chances of detecting phishing websites, which can in turn help to tip the balance back away from the attacker.

### B. How users respond – human behaviour analysis

The fact that phishing websites are specifically designed to deceive users and exploit their lack of attention and security knowledge stems from the attackers' understanding of how users actually behave when interacting with websites. Indeed, as Adams and Sasse state: "hackers pay more attention to the human link in the security chain than security designers do" [6]. To redress this balance, researchers in the field such as Kirlappos and Sasse [13] suggested that it would be helpful to consider the entire decision-making lifecycle, from the point at which the user perceives and starts to pay attention to the website, through to their final decision as to whether or not to trust the site with their information.

*1) Perception and attention:* When a user first views a website there are two principal, and different, processes at work: perception and attention. The user will initially perceive all the objects on the site (text, logos, images, etc.) within their field of vision, and will take in certain basic information – the user will not however really be looking at the site until he pays attention to a given object within this field of vision [8]. The challenge faced by users is that their attention is typically drawn to objects with highly vivid, salient features, and away from less salient, but more important security indicators elsewhere on the webpage. If this tendency is not overridden by sound security knowledge, users will be likely to fail to notice vital clues as to the validity or otherwise of the website. This is neatly illustrated in [14], with eye movement studies highlighting how expert users paid attention to objects of high security value such as an SSL padlock icon in the address bar, while novice users had their attention drawn instead to attractive, but irrelevant, logos in the content area of the page.

*2) Knowledge and learning:* Having paid attention to the objects on the page, the user must then interpret them within the context of the core decision to be made: "is this site genuine?". Whether the user correctly interprets the objects in this context depends upon what they know and have learned about their security values. Cranor again [15] made the point that security indicators are of limited use if users do not understand what they mean.

And yet many users have low levels of security knowledge, failing to understand the role of the HTTPS padlock, or how to interpret a URL. Some of this confusion stems from a more general lack of IT and internet knowledge [12], while some is due to the complexity of the topic – URLs are not straightforward to understand. Many argue, however, that another root cause is the low quality and effectiveness of security education generally provided to users.

*3) Memory:* Of course, even if users possess website security knowledge this is of limited value if they cannot remember it at the point of need. Psychologists often consider two types of memory – working memory, which is of limited capacity and which decays away within a few seconds, and long-term memory, which does not suffer from capacity issues but which cannot be retrieved easily [16]. The ability to retrieve and access knowledge from long-term memory is dependent upon factors such as how frequently the knowledge

is used, how recently it has been retrieved, and how the user was impacted by the use of this knowledge. Individuals who regularly practice applying their security knowledge are therefore more likely to be able to recall critical security knowledge than those who do so less frequently.

*4) External factors*: Human responses in the face of a phishing attack can also be affected by a range of external factors, such as social engineering techniques [17] which seek to exploit the fact that humans typically make more errors when placed under time or social pressure, with a common example being the perceived need to enter credit card data quickly in order to acquire tickets online before an offer period expires. These factors are effective as they lead to an increase in cognitive workload which can in turn affect the user's quality of judgement as to the site's authenticity.

*5) Risk-based decision making:* The above factors – perception of and attention to the website, the learning and retrieval of website security knowledge, and external factors such as time and social pressure – all combine to enable the user to make a judgement as to the level of trust which he or she feels can be placed in the site. The actual decision as to whether to enter sensitive information into the site, however, is a risk-based decision, in which the user has to determine whether the overall risk of providing the information is outweighed by the benefit of doing so. The risk component involves the traditional calculation Risk = Probability of Occurrence × Impact, which in turn requires an assessment of asset value (how sensitive is my information, and how useful is it to others?), the perceived level of threat to those assets, the levels of vulnerability within the site, and the potential cost to the user should the risk materialise. The benefit side to the equation, meanwhile, involves an assessment as to the potential gain associated with submitting information to the site. It is worth here reiterating the point made by Anderson [7] and Moore and Clayton [18] that users are not purely driven by security considerations, but by reward, and hence their decision-making process will take this into account.

*6) How people think – 2 systems:* Despite what the above discussion may imply, user behaviour is of course an immensely complex process which does not even conform to a single "system" concept. Kahnemann [19] has argued that the brain operates 2 systems: System 1 aligns with our idea of "gut instinct", making automatic, rapid, and effortless decisions and judgements based on educated guesses, rules of thumb, pattern recognition and heuristics. System 2, meanwhile, is described as "deliberate", "effortful", "slow", and based on reason – this is the system which we might recognise as we find ourselves actually thinking hard about a problem, for example. Intuitively we can understand that both systems are used during the process of determining whether to trust a website, but which system is used at what stage and under what conditions is still far from being well understood.

### C. *Modelling human behaviour – the ACT-R cognitive architecture*

The ACT-R cognitive architecture [20] is both a framework for human cognition and a model which is implemented in software using Lisp. The architecture reflects the modular structure of the brain, and represents a huge body of research into psychology and cognitive science.

The architecture sets out, as shown in Fig. 1, a series of modules which relate to different regions of the brain; buffers through which the modules interact and exchange information; and productions which contain the rules which govern how the modules interact. These modules perform several roles: visual and motor modules simulate the brain's interface with the external world; declarative and procedural memory modules deal with factual and procedural (rule-based) knowledge and its storage in memory; while the goal module keeps track of the end objective which the model is trying to achieve.

The final component within the ACT-R model is its pattern-matching capability, which can change the state of the system by activating, or "firing", different productions, which in turn apply rules to change the state of individual modules by modifying their buffers. In this way, the model can advance through a task from, for example, visual interpretation of a website through to manual confirmation of the decision as to the website's validity.

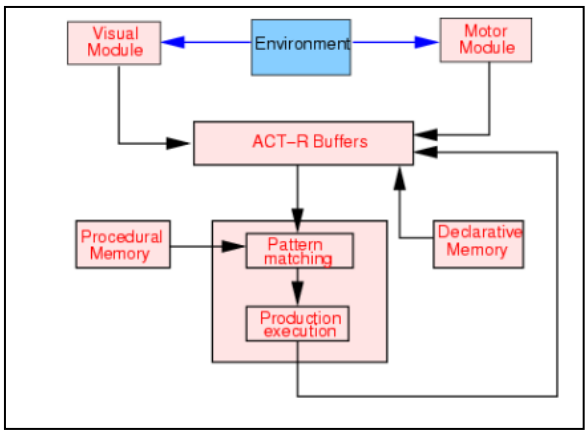


Fig. 1. How ACT-R works – illustrative example

## III. MODEL AND EXPERIMENT DESIGN

### A. *Conceptual model*

To bring together the range of factors which influence human behaviour and decision-making in the face of a phishing attack, a high-level conceptual model was first developed. This model, whose scope was limited to the process of deciding on the validity of a given webpage, also served as an overarching framework for the more detailed modelling of the process within the ACT-R architecture, ensuring that the simulations of human behaviour were aligned to cognitive behaviour theory.

The inputs to the model consist of initial base information, which enables core human attributes such as knowledge levels and risk appetites to be represented, and the representation of the candidate website and goal setting for the experiment,

namely "is the website genuine or fake?". Given these inputs, the model then applies the core cognitive processes to arrive at a decision as to the trustworthiness of the site, before finally outputting the decision, which can then be compared with the actual status of the candidate website under consideration.

This model, while prioritising the human cognitive behaviour processes, can also be seen to encapsulate the principal concepts and components of the security risk assessment process: valuation of the information assets at risk; assessment of that risk based on perceived threat levels; the phishing attack itself; the vulnerability of the user to the attack; and the controls in place to mitigate the risk, such as user training or the inclusion of the HTTPS padlock. These factors all combine to simulate the user making an overall residual risk calculation based on the probability of an attack occurring and the likely impact to the user in the event the attack does occur.

### *B. ACT-R model design: HTTPS padlock*

This high-level logical model provides an appreciation of the various factors which influence a user's decision as to whether to trust a website, but is purely theoretical in nature, and hence cannot be used to test the accuracy with which it reflects human behaviour; nor is it sufficiently detailed to predict how humans may respond, or how effectively, when confronted with a potential phishing website. It is therefore necessary to model the processes and interactions relating to human phishing website detection within software, and then define and run a series of experiments through the model whose results can be compared to those of actual human trials.

As preliminary work to prove the concept, we decided to focus specifically upon modelling a basic website representation with a primary focus on the role of the HTTPS padlock. This selection supported the primary aims of the study since the padlock is present on all genuine HTTPS webpages (and not, generally, in fake ones) and hence can be used as a valuable indicator of webpage validity. It is also something which many users struggle to understand, verify, or even recognise, enabling the impact of user knowledge on the webpage validity decision to be investigated. The padlock's key characteristics (presence/absence; location; colour; size; verifiability) are also relatively simple to model. Finally, some baseline reference data [14] was available to enable the model's performance to be compared against some benchmark.

The ACT-R HTTPS padlock sub-model design, as shown in Fig. 2, was developed to reflect the key conceptual model characteristics:

- **Webpage representation**, in which a representation of a webpage is generated on a screen, upon which objects representing a security padlock icon and other, non-security icons are displayed
- **Determination of webpage characteristics**, which enable a simulated user to characterise the webpage as genuine or fake based on the presence or otherwise of the representative padlock icon, the visual appearance of the icon, and the location of the icon on the screen
- **Pre-defined user knowledge levels**, including the ability to draw upon and utilise different facts and rules relating to webpage security depending on the level of security knowledge of the end-user being modelled
- **Perception of the webpage** and the objects on the screen, together with an ability to focus on ("attend to") certain objects on the page
- **Access to, and storage of information in memory:** specifically, the model's ability to encode and store information perceived on the screen as well as retrieve factual and procedural (rule-based) security knowledge
- **Decision-making**: the model design, to represent the human cognitive process, required the ability to apply procedural and factual knowledge using a pattern-matching, rather than sequential, approach in deciding whether the candidate webpage is genuine or fake
- **Randomness**, introduced to vary the selection and placement of the security padlock icon and other, non-security, icons to prevent the model becoming deterministic and hence predictable in its outputs.

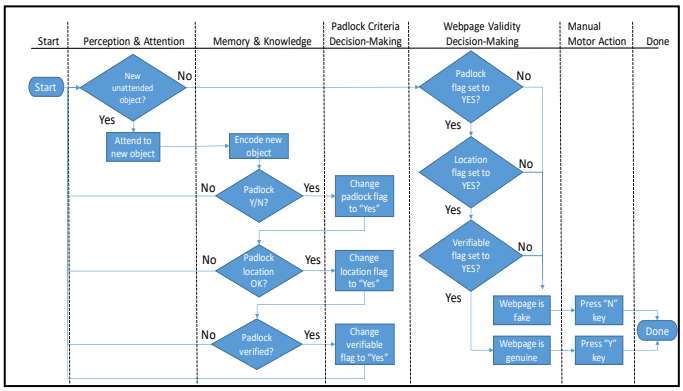


Fig. 2. ACT-R padlock model design overview – expert user

The decision-making requirement in particular reflects the need for the model to follow the ACT-R approach rather than that traditionally adopted computer model design. As a representation of the human mind, ACT-R is a complex yet powerful system – it has the ability to learn and remember facts, rules and strategies, and has been used in the study of a range of tasks covering perception and attention, language and communication and problem-solving and decision-making [21]. ACT-R is not, however, a linear tool and hence its instantiation in software is therefore noticeably different from more traditional software implementations. The modeler must therefore take this into account, set up the rules which ACT-R must obey and the conditions within which it must operate, and then let the model itself dictate the sequence of events.

### *C. Experiment design*

Once the core model had been designed, a set of experiments were then needed to be defined to test the model's effectiveness. Key requirements for the experiments included: the ability to vary the input parameters provided to the model during different experiments (for example to test the model's effectiveness against different webpage representations or levels of user knowledge); the ability to compare the model's output decision with the actual webpage validity status under

these different conditions; and the potential for future comparison between the model's performance and that of actual end users when presented with the same conditions.

To test the effectiveness of the model from a proof of concept perspective, while also seeking to simulate as realistically as possible an end user's interaction with a suspected phishing website, three core experiments were considered, each building on the aims of the previous one. In each case, and for each trial run, a start state for the model was defined, setting initial conditions such as the level of user knowledge, the representation of the candidate webpage, and the model's initial judgement to the webpage's validity (set to "unknown"). The aim of each experiment was then to present the candidate webpage to the model, then set the model the task of judging whether the website was genuine or fake.

*1) Experiment 1: Phishing website detection performance – novice user:* The aim of this experiment was simply to investigate how well the padlock model, together with a simulation of a novice user, would be able to represent a real end-user faced with a potential phishing website. For this experiment, the user knowledge level was set to "low".

*2) Experiment 2: Low knowledge versus high knowledge:* The aim of the second experiment was to introduce users with higher levels of security knowledge and then compare their performance with those of the novice users in the expectation that more highly skilled and knowledgeable users would perform better than novice users. This experiment sought both to determine the extent to which this might be the case, and to understand the factors which drive the model's performance.

*3) Experiment 3: Effects of randomness:* The aim of the third and final experiment was to test how the quality of user decision-making was affected by variations to the simulated user's ability to recall (i) key factual security knowledge relating to the HTTPS padlock, and (ii) key rule-based security knowledge which guides the process of determining the webpage's validity. In particular, the experiment sought to understand whether the role of randomness within the decision-making process (reflecting user distraction or forgetfulness, for example) could improve the realism of the model in simulating actual human behaviour.

## IV. IMPLEMENTATION, RESULTS AND ANALYSIS

### A. Core model and experiment development and testing

Three versions of the model were developed during this study. The first version sought to demonstrate the core ability of the program to perceive a series of objects placed on a simulated webpage, to "test" the security attributes of these objects according to a set of rules contained within the model relating to the characteristics of the padlock, and to output a decision as to whether the webpage was genuine or fake.

The first step was to build a basic representation of a webpage using the ACT-R window (an in-built function within ACT-R), onto which a series of textual objects were added as shown in Fig. 3. Some of these objects were placed at fixed locations, such as the URL, name of an imaginary bank owning the webpage, and a solid line indicating separation between the address bar and content Areas Of Interest (AOIs) of the webpage. Other objects, notably the single-letter objects, were placed randomly on the screen and subject to certain constraints, such as the placement of one of the (randomly selected) objects within the security icon AOI on the left-hand side of the address bar.


Fig. 3. ACT-R 'webpage' with address bar security indicator representations

Next, the base level of user knowledge was encoded. In this first version of the model, a low level of knowledge was assumed, so only basic (and incomplete, and sometimes even incorrect) facts regarding webpage security were provided to the model, namely: factual knowledge as to what an HTTPS padlock looks like (represented by a letter "X" in this model); the ability to recognise a padlock on the screen; and a rule-based item of knowledge which simply stated that if a padlock is seen in the context AOI of the screen then the webpage is to be considered genuine and can be trusted.

Note here that in the case of a novice user, the significance of the location of the padlock on the screen is not appreciated, reflecting the tendency for novice users to be attracted to logos suggesting website security even when such logos offer no security assurance.

Once these initial model inputs had been encoded, the core cognitive processes were modelled. This was achieved by encoding a number of ACT-R procedures, each focusing on a different cognitive process, whether this be attending to an object perceived on the screen, storing new knowledge for future retrieval or use, or recalling factual or rule-based knowledge from memory in order to make a decision.

The last component of the core cognitive model to be encoded was the command to the ACT-R manual model to confirm the decision by "pressing" a key to record the judgement as to whether the webpage is genuine (indicated by pressing the letter "Y") or fake (the letter "N"). With this output the model is deemed to have completed its tasks, and hence the goal state is updated to "done". At this point the experiment once again takes control, comparing the model output to the true status of the webpage to determine whether the model was successful in its judgement.

With the coding of the model and the experiment completed, the following four scenarios were then defined to test both the ability of the model to function as expected and its performance in judging the webpage's validity.

- Scenario 1: user is presented with a genuine webpage ("X" in the security icon AOI) in which a padlock logo ("X") is also displayed in the content AOI
- Scenario 2: user is presented with a genuine webpage ("X" in the security icon AOI) but no padlock logo is displayed in the content AOI
- Scenario 3: user is presented with a phishing webpage (no "X" in the security icon AOI) but one or more padlock logos are displayed in the content AOI
- Scenario 4: user is presented with a phishing webpage (no "X" in the security icon AOI) and no padlock logo is displayed in the content AOI

The experiment was then run multiple times, with a scenario selected randomly on each occasion, and the relevant webpage generated and presented to the model for analysis.

The results of this initial experiment showed the model's overall performance (phishing webpage detection success) to be low, with a success rate of only 38% in correctly determining the validity of the candidate webpage. Looking in more detail at the model's performance in each of the above scenarios, it becomes clear that the (simulated) novice user's inability to pay any attention to whether a padlock icon is displayed in the security icon AOI results in a decision-making process which is no better than guesswork. For example, a user will consider the webpage to be genuine if an "X" is perceived within the content AOI – this decision is correct if there also happens to be an "X" in the security icon AOI of the screen, but the correct decision is based on chance and luck rather than the application of security knowledge. Similarly, a user may correctly determine the webpage to be fake if he does not perceive a security logo in the content AOI and there is no security padlock icon within the security icon AOI.

It was noted that the performance figures above were significantly lower even than the findings of the eye tracker experiment [14], which reported an average error rate of 32.4% among novice users attempting to determine website validity. Direct comparison between these figures is inappropriate, however, due to the differing conditions under which the experiments were run, not least the fact that the eye tracking experiments were based on presentation of 12 phishing websites and 8 genuine websites as opposed to the ACT-R model (16 phishing websites, 4 genuine). It would also be unwise, given the rather crude nature of this version of the model, to place too much importance on the above performance figures. Nevertheless, the results would seem to indicate that this version of the ACT-R model was, to some degree at least, broadly performing in a similar way to novice users when confronted with a potentially fake webpage and asked to make decisions based on little/no security knowledge.

### *B. Improved search strategy: novice and expert users*

The principal aim of the second version of the model was to refine the initial program to better reflect the decision-making strategy of end users possessing greater knowledge of website security, and specifically those with knowledge of the security attributes relating to the SSL padlock and URL.

Unlike novice end users, more expert users will typically focus on the address bar, rather than the content AOI, within a webpage to seek indicators relating to the webpage's security. The productions within the model were therefore amended to reflect this improved factual and procedural knowledge.

Initial testing of the revised model did, as expected, simulate the behaviour of advanced users focusing on the address bar AOI while determining whether the webpage was genuine, and on the basis of 100 trials the model achieved an overall success rate of 64%. On the occasions that the genuine padlock was displayed on the screen and the model focused on this object, the model correctly determined that the webpage was genuine. The model did not, however, focus on the padlock AOI in all cases, while in other trials the padlock was not present (indicating a fake website), and yet the model failed to correctly identify the webpage as such.

Analysis of these results showed that running the experiment against the more knowledgeable "intermediate" model demonstrated marked improvement in the success rate compared with the novice user's performance in version 1. Again, this is to be expected, since this model relies on the application of actual security knowledge rather than the novice user model's mistaken belief in the value of security logos placed within the content AOI of the page. Interestingly, despite the absence of specifically defined randomness or "noise" within the model's functionality, the model itself did not predict the webpage validity status accurately 100% of the time. Failures to detect genuine websites were attributed to the model perceiving other objects within the address bar AOI and therefore, like the novice model, basing its decision on irrelevant information.

This version of the model however, despite simulating advanced user behaviour much more faithfully than version 1, still suffered from a number of limitations which prevented it from fully reflecting the way in which an expert user would apply his security knowledge to a webpage. Most significantly, an expert user would (unless highly distracted), be unlikely to completely ignore the presence or otherwise of a security logo in the security icon AOI of the address bar. In addition, an expert user would have applied his knowledge of genuine URLs to the URL seen on the screen in deciding upon how much to trust the webpage – while it is not known what weighting such users would typically apply to this decision, it is reasonable to assume that security indicators contained within the URL itself do carry some weighting, and hence the absence of this capability from the model is a limitation. It is therefore suggested that URL analysis be considered as a future enhancement to the model, and that this as a minimum should include analysis of the protocol (http / https) used.

### *C. Introduction of probabilistic decision-making*

In reality of course, judgements as to a webpage's validity are never based on 100% certainty but are rather risk-based in nature, with a user deciding that a given webpage is "sufficiently likely to be secure" to trust it with sensitive information. The aim of version 3 of the model was therefore to introduce an element of probability, and hence unpredictability, into how the simulated human would behave,

so that the model can more accurately reflect, and predict, actual human performance.

As discussed above, there are many factors in addition to the level of security knowledge which influence the decision-making process. Critical to an end user's ability to correctly identify a website as genuine or fake however is his ability to *remember* what he has learned about website security, and his ability to *avoid distractions* during the decision-making process. It was therefore decided to introduce both these factors into the model, utilising the probabilistic functionality within ACT-R to reflect the possibility that a user may on occasions either forget previously learned factual security knowledge or else fail to apply security rule-based knowledge correctly.

*1) Imperfect recall of factual knowledge:* Within this version an element of distraction was introduced, so that the model would, on occasions, fail to "remember" the significance of the padlock when perceived on the screen. The degree to which the user is "distracted" was adjusted within ACT-R by varying the level of noise presented, with greater noise levels increeasing the likelihood that the fact would not be retrieved from memory. Analysis of the model's performance with varying levels of noise revealed a slight downward trend as noise levels increased. This result aligns with expectations, with the simulated end user making more mistakes as the level of distraction increases. With very high levels of noise, however, the model continued to perform reasonably well, achieving a 70% success rate, suggesting that noise levels only have a certain level of impact on the quality of the model's decision-making, and that beyond this level other factors become more significant.

*2) Imperfect application of rule-based knowledge:* Rule-based knowledge is applied within ACT-R through the use of the productions, which become candidates for activation, or "firing", if the test conditions within the production are satisfied. If multiple productions meet a given set of test criteria, then ACT-R will select one of these productions to fire based on their utility – a parameter which each production possesses and which is used by ACT-R during conflict resolution between productions [22]. Version 3 of the model used this functionality to introduce a small element of probability that expert users would look in the content AOI for security indicators rather than the address bar AOI. The introduction of this imperfect application of rule-based knowledge was expected to result in a slight reduction in the expert user model's effectiveness, with its performance decreasing steadily with the corresponding decrease in correct application of the rule. On the occasions that the trial included a large degree of noise this result was achieved, successfully demonstrating the concept of user perception and attention being affected by noise in the system "distracting" the user from applying the rule-based security knowledge which he would apply had the distractions not been present. However. this behaviour within the model was displayed at a cost, namely a significant increase in time spent by the model in attempting to reach its decision, and a significant reduction in the model's overall performance, which was assessed to be associated with the noise affecting the whole model, and not simply those target productions.

## V. DISCUSSIONS

### A. Applicability to the phishing attack and detection process

The ACT-R architecture, and its instantiation in software, delivers significant capabilities in the realm of human cognitive behaviour modelling which are highly applicable to the phishing website context, and indeed across the full phishing attack lifecycle. The core ACT-R functionality used in the HTTPS padlock model for example – perception and attention, knowledge and memory, problem-solving and decision-making, and motor-based confirmation of decision – delivered a simulation of human behaviour which was fully recognisable as reflecting the way in which both novice and expert users would behave throughout a security scenario.

### B. Cognitive modelling capabilities

While some of the core cognitive modelling capabilities within ACT-R were successfully applied to the proof of concept phishing scenario within this study, a review of the ACT-R reference manual [22] also reveals a wide range of capabilities which were unused in the specific design implementation but which would deliver great value if incorporated within future models of human phishing detection – ACT-R's ability to learn, for example, could enable simulation of a process by which users are trained to recognise phishing websites and receive feedback on their performance to improve their overall ability.

### C. User interface / interoperability

Less positively, and because of its low-level nature and basic user interface, it would not appear to be a straightforward task to provide an actual webpage as input to the model. The inability of the ACT-R visual module to perceive and attend to objects within an actual webpage rather than a crude representation of a webpage was an unwelcome limitation of the overall model.

### D. URL modelling

It is also considered that the model would have been greatly enhanced had it included not only the HTTPS padlock functionality but also a more realistic simulation of user interaction with URLs. A high-level design for this component of the model was produced, and a sample URL included within the representative webpage with the components of the URL (protocol, domain name, parameter list) represented as different objects within ACT-R. However, it became apparent that the core model was unable to interpret the URL as separate objects, instead focusing on individual letters or symbols in a way which was not representative of human behaviour. The model's functionality did not extend to interpretation and application of knowledge to the objects within the URL as perceived by the model, and hence the overall effectiveness of the model was limited. Incorporation of robust URL perception, knowledge/memory retrieval and problem-solving functionality within the model would be a valuable piece of

future research, since this could be combined with the model's analysis of the HTTPS padlock to deliver significantly more advanced decision-making capabilities.

### E. *User benchmarking*

Finally, although not a limitation of the ACT-R model, the failure to identify any research reporting the results of user phishing detection trials which mirrored the experiments within this study meant that it was not possible to draw any meaningful conclusions regarding the model's accuracy. The results from the study did broadly align with what might be expected from actual end users, and the priority for the study was to develop a proof of concept model rather than develop highly accurate predictions of actual end user behaviour. Nevertheless, further refinement of the model and greater experimentation to enable further analysis of model performance would have been of value.

## VI. Conclusions

This study has investigated how one of the most complex parts of the security system – the human – interacts with one of the most prevalent and effective forms of attack – phishing. In doing so it has identified a significant opportunity to build on previous research in this space to gain still greater insights through the application of the ACT-R cognitive architecture. The study has found that despite the low-level nature of the architecture and basic user interface, ACT-R possesses strong capabilities which map well onto the phishing use case, and that future research could usefully build on this initial proof of concept model. Suggested lines of development include: refining the current "proof of concept" HTTPS padlock sub-model to better exploit ACT-R's functional capabilities and improve overall model performance; improving the interface between ACT-R and websites to enable actual websites to be presented to the model; extending current model capabilities by incorporating additional security indicators and simulating additional user behaviours; comparing the ACT-R model's performance with real user trials; and expanding the scope of the model to encompass the entire phishing attack cycle.

## Acknowledgements

Nick Williams thanks Dr Adrian Banks and Dr Patrice Rusconi from the School of Psychology, University of Surrey for generously giving their time and expertise, both in helping the challenges within cyber security to be viewed from a new perspective, and in framing the investigation into phishing attack detection as a psychological as well as a security question. Shujun Li was partly supported by the UK part of a joint Singapore-UK research project "COMMANDO-HUMANS: COMputational Modelling and Automatic Non-intrusive Detection Of HUMan behAviour based iNSecurity'", funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/N020111/1.

## References

[1] CERT-UK (now part of National Cyber Security Centre – NCSC), "Phishing: What is it and how does it affect me?" 2015. [Online]. Available: https://www.ncsc.gov.uk/guidance/phishing-what-it-and-how-does-it-affect-me. [Accessed: 15-May-2017]

[2] C. Green, "Top ten things you need to know about data breaches," Information Age, 2015. [Online]. Available: http://www.informationage.com/technology/security/123460135/top-ten-things-you-need-know-about-databreaches. [Accessed: 15 May 2017]

[3] APWG, Phishing Attack Trends Report – 1Q 2016, 2016. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q1_2016.pdf. [Accessed: 15 May 2017]

[4] B. Schneier, "People, process, and technology," Schneier on Security, 2013. [Online]. Available: https://www.schneier.com/blog/archives/2013/01/people_process.html. [Accessed: 15 May 2017]

[5] ISO/IEC, Information technology – Security techniques – Information security management systems Requirements, ISO/IEC 27001:2013, 2013

[6] A. Adams, "Users are not the enemy," *Communications of the ACM*, vol. 42, no. 12, pp. 41–46, 1999

[7] R. Anderson and T. Moore, "Information security: where computer science, economics and psychology meet," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1898, pp. 2717–2727, 2009

[8] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychological Review*, vol. 111, no. 4, pp. 1036–60, 2004

[9] ACT-R Research Group, ACT-R Software. [Online]. Available: http://act-r.psy.cmu.edu/software/. [Accessed: 15 May 2017]

[10] C. Lebiere, S. Bennati, R. Thomson, P. Shakarian, and E. Nunes, "Functional cognitive models of malware identification," *Proceedings of 13th Annual International Conference on Cognitive Modeling*, pp. 90–95, 2015

[11] B. Schneier, "The security mindset," Schneier on Security, 2008. [Online]. Available: https://www.schneier.com/blog/archives/2008/03/the_security_mi_1.html. [Accessed: 15 May 2017]

[12] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," *Proceedings of 2006 SIGCHI Conference on Human Factors in Computing Systems*, pp. 581–590, 2006

[13] I. Kirlappos and M. A. Sasse, "Security education against phishing: A modest proposal for a major rethink," *IEEE Security & Privivacy*, vol. 10, no. 2, pp. 24–32, 2012

[14] D. Miyamoto, G. Blanc, and Y. Kadobayashi, "Eye can tell: On the correlation between eye movement and phishing identification," *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings Part III, Lecture Notes in Computer Science*, vol. 7666, pp. 223–232, 2015

[15] L. F. Cranor, "What do they 'indicate?'," *Interactions*, vol. 13, no. 3, pp. 45–47, 2006

[16] N. Cowan, "What are the differences between long-term, short-term, and working memory?" *Progress in Brain Research*, vol. 169, pp. 323–338, 2008

[17] K. D. Mitnick and W. L. Simon, *The Art of Deception: Controlling the Human Element in Security*, John Wiley & Sons, Inc., 2002

[18] T. Moore and R. Clayton, "The impact of public information on phishing attack and defense," *Communications and Strategies*, no. 81, pp. 45–68, 2011

[19] D. Kahneman, Thinking, Fast and Slow, Penguin, 2012

[20] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, 2007

[21] ACT-R Research Group, "ACT-R." [Online]. Available: http://act-r.psy.cmu.edu/. [Accessed: 15 May 2017]

[22] D. Bothell, *ACT-R 7 Reference Manual*, 2015. [Online]. Available: http://act-r.psy.cmu.edu/wordpress/wp-content/themes/ACT-R/actr7/reference-manual.pdf. [Accessed: 15 May 2017]