Mixed-Initiative Security Agents

Rachel Greenstadt
Drexel University
Philadelphia, Pennsylvania
greenie@cs.drexel.edu

Sadia Afroz
Drexel University
Philadelphia, Pennsylvania
sa499@cs.drexel.edu

Michael Brennan
Drexel University
Philadelphia, Pennsylvania
mb553@cs.drexel.edu

ABSTRACT

Security decision-making is hard for both humans and machines. This is because security decisions are context-dependent, require highly dynamic, specialized knowledge, and require complex risk analysis. Multiple user studies show that humans have difficulty making these decisions, due to insufficient information and bounded rationality. However, current automated solutions are often too rigid to adequately address the problem and leave their users more confused and inept when they fail. A mixed-initiative approach, in which users and machines collaborate to make security decisions and make use of complementary strengths rather than weaknesses, is needed.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous; K.4.2 [**Computers and Society**]: Social Issues —*Abuse and crime involving computers*; K.6.5 [**Security and Protection**]: [Authentication,Unauthorized access (e.g., hacking, phreaking)]

General Terms

Security,Human Factors

Keywords

Mixed initiative, Artificial Intelligence, Security

1. INTRODUCTION

Techniques from artificial intelligence (notably bayesian learning and captchas) have achieved great success in helping administrators manage automated attacks such as SPAM and network attacks that would overwhelm human capacities [21, 18]. This paper argues, however, that artificial intelligence techniques have an even greater role to play in the security story.

Security decision-making is hard for both humans and machines. Security decisions are context-dependent, require specialized knowledge, are highly dynamic due to sophisticated adversaries and evolving threats, and require complex risk analysis. Multiple studies show that humans have cognitive difficulties making these decisions [1, 6, 9, 19, 20, 22]. Users either retreat from online activities (causing lost opportunities) or throw caution to the winds. This results in expenses incurred cleaning up after infections. These users unwittingly support the computer crime infrastructure with their computational and financial resources.

Current automated solutions are often too rigid to adequately address this problem, and leave their users more confused and inept when they fail [8]. Automation annoys users when it prevents them from getting their primary work done [1].

A **Mixed-initiative approach** refers broadly to methods that explicitly support an efficient, natural interleaving of contributions by users and automated services aimed at converging on solutions to problems [14]. The term comes out of the user interface community and reflects a rejection of the choice between interfaces that allow users to directly manipulate objects and interfaces that sense user activity and take automated actions. We argue that a mixed-initiative approach to security decision-making is needed, in which users and machines will collaborate to make security decisions, making use of complementary strengths rather than weaknesses. Such an approach will require shared representations of contextual information, well-designed interfaces, adversarially-resistant learning mechanisms, and trustworthy methods for incorporating global information from outside sources.

This paper extends the ideas of adjustably autonomous security agents [11]. A potential design for such an agent is shown in Figure 1. A discussion of how virtualization can be used to protect security agents from compromise can be found in [11].

2. DESIGN OF SECURITY ASSISTANT AGENTS

Recent research has shown that humans have difficulty in consistently performing repetitive tasks, memorizing large amounts of information, and managing the accumulation of small risks or understanding long-term risk [20]. However, machines lack common sense reasoning and often fall behind humans in strategic planning tasks. They can be fully compromised by control hijacking attacks. Finally, there will always be cases when machine automation fails and humans must be relied upon. Currently, computer security often falls in a worst of all worlds in which attackers are able to exploit both machine and human weaknesses.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
AISec'09, November 9, 2009, Chicago, Illinois, USA.
Copyright 2009 ACM 978-1-60558-781-3/09/11 ...$10.00.

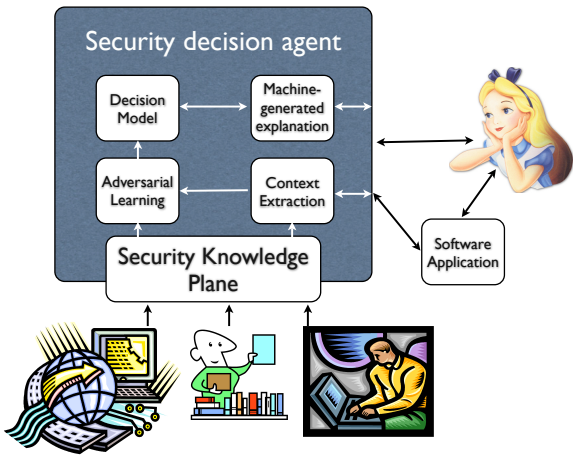

Figure 1: Proposed design of a security decision agent. The agent interacts directly with user Alice and her software application and can also draw on knowledge-bases on the Internet, contributed by experts and other users.

Decision Models. However, we can and should deconstruct security decision-making to understand which components of the decision are best made by humans and which are best made by machines. Once this is better understood, we can build security assistant agents, where humans and machines can complement each others' expertise when making these decisions.

Context extraction. When programs ask their human users to make decisions about security, the reason that the decision cannot be fully automated is often due to the lack of contextual information that the program has about the decision. Security decisions often rely on several contextual factors for input including (1) the software and network activities going on concurrently with the decisions, (2) knowledge about the resource being accessed (the webmasters, code authors, or certificate authorities), and (3) the beliefs, desires, and intentions of the human users interacting with the system. When security assistant agents can gather more of this contextual information, we hypothesize that they can much better assist their users in making security decisions.

Machine generated explanations of security risk factors/decisions. If humans and machines are to collaborate in decision-making processes, then the reasoning behind computer input to these decisions must not be opaque to humans. Certain machine learning representations, such as decision trees, are more understandable to humans than other methods, such as SVNs or neural networks. When security agents make recommendations, we need to find ways to translate the reasoning behind these decisions to something humans (and particularly humans without strong technical or security backgrounds) can understand and assimilate into their view of the world. Good work along these lines has been done in manipulating nuanced privacy preferences [17].

Further challenges include adapting AI techniques to adversarial situations such as **adversarial learning** and building a **knowledge plane** that incorporates potentially unreliable or adversarial information from other humans and/or software agents.

The ultimate goal is an agent system like that illustrated in Figure 1. While this may sound daunting we show in the following subsections that minimal amounts of context extraction and understanding of adversarial learning capabilities can make a profound impact on security decisions such as "Should I trust this website" or "Is this piece of writing likely to reveal my identity?"

2.1 Shared representations and phishing detection

Phishing is a web-based attack that uses social engineering techniques to exploit Internet users and acquire sensitive data. According to the Anti-Phishing Working Group (APWG), there were at least 47,324 phishing attacks in the first half of 2008 [12]. Most phishing attacks work by creating a fake version of the real site's web interface to gain the user's trust. Despite many browser-based indications, user studies have shown that over 90% users trust any site based on its appearance, just as they use the appearance of physical sites to judge their authenticity [10, 7].

We argue that effective phishing detection mechanisms must detect phishing sites from the user's point of view. That is, the detection should be directly related to the look and feel of the site as most users take security decision based on sites' look (for example, the logo and page layout).

Most effective phishing detection methods depend on blacklisting. Blacklisting is mostly user dependent and relies on manual verification. As most phishing sites are too short-lived to update and verify the database, the blacklisting approach fails to detect many phishing attacks. Furthermore, a blacklisting approach will fail to detect an attack that is targeted to a particular user ("spearphishing"), particularly those that target lucrative but not widely used sites such as company intranets, small brokerages, etc.

The majority of users provide sensitive credentials to a small set of sites (fewer than 20) [4]. Under the assumption that SSL is supported by these sites of interest and secure in both the underlying protocol and the trust model used by the browser, these sites can be whitelisted and browsers can automatically verify their authenticity. The problem with whitelisting approaches [13, 5, 15], is that the user must know about and remember to check the interface every time they visit the site, and there is ample evidence that this is beyond most users' capacities. However, warning users when the site they are visiting is not among their sensitive subset is also futile, as the vast majority of sites visited by users are not sensitive and such warnings will be quickly tuned out or turned off. What is needed is for the browser to infer the user's *false belief* that she is visiting one of her sensitive sites and only warn (actively and emphatically) in this case. Our hypothesis is that similar-looking content can be detected by automated methods. The small set of sites that an individual user use can easily be rendered and saved as that user's personal profile set. Phishing sites that imitate any site in that profile set can be detected by matching contents (e.g. URL, images, most used texts, HTML codes, scripts, etc.), which are related to its structure and appearance, with those of the real site. Preliminary results suggest that most phishing sites use copied contents from the corresponding real sites and—by using fuzzy matching—current attacks (as measured by mining www.phishtank.com can be

detected with 97% accuracy and high precision (less than 1% false positives) [2].

The goal of the attacker will be to make websites that are different to computer algorithms, but (close to) identical to human eyes. If the attackers still prove successful in defeating our matching algorithms, they will have contributed to our understanding of the vision problem in much the same way that spammers have improved statistical machine learning and bots that defeat captchas have improved optical character recognition algorithms. If these techniques succeed, but reduce the efficiency of our techniques, they can be run offline (on email links) or by intermediaries.

We argue the reason that this approach works well is that it creates a shared representation between the user and the browser in the form of a content profile for each trusted site. This allows the browser to recognize imitations by content matching and helps user to make security decisions by providing active warnings.

2.2 Adversarial learning and anonymous publishing

The web is full of anonymous communication that was never meant to be analyzed by authorship recognition systems. An anonymous message board, for example, is often not meant to reveal which posts are by which authors, or how many authors exist on the forum in the first place. While posters can hide their IP addresses using anonymous communication protocols such as Tor, the linguistic content of their posts might still give them away.

For this reason, it is important to understand the degree to which machine learning-based authorship recognition techniques are effective. We have found that three authorship recognition techniques, which are representative of current trends in the field, fail when ordinary users try to hide their writing style [3], even though are quite effective when users do not modify their writing. Other recent research has shown that text analysis can be used to find and modify the most salient features in a document in order to protect the anonymity of the author [16].

The results of these studies demonstrate that (a) it is possible to retain privacy against current stylometric techniques, (b) the high effectiveness of authorship techniques on unmodified documents suggests it important for users who desire privacy to take measures to hide their identity, and (c) that automation can be used in order to detect the best means for obfuscating a document.

While it may be easy for an algorithm to modify a document in order to preserve anonymity, it is a much more complicated task to do so in such a way that preserves the semantic content of the text. And while a human can better modify a document without obscuring it's meaning, it is unreasonable to expect a person to perform complex analysis on the text to analyze the most vulnerable features. This paves a clear path for building an agent that assists users in determining when it would be worthwhile to obfuscate their writing and the most effective ways of doing so.

3. CONCLUSION

This position paper argues for the development of a mixed-initiative approach to security, in which users and machines collaborate to make security decisions and make use of complementary strengths rather than weaknesses.

If a scientific foundation for mixed-initiative security agents can be successfully developed, then integrated into browsers, operating systems, and applications, it will make the work of the attacker much harder. Currently, there is a choice for people between participation in Internet life and risk. Those who are less educated and computer savvy face larger risks and often preyed upon by identity thieves, scammers, and other attackers. They are used to build the infrastructure (botnets) to attack more hardened targets. Improving security decision-making at the end user level can have a broad impact on overall computer security.

Examining these ideas will help us to understand the relationship between the fields of HCI, AI, and security to the benefit of all three. Security decisions provide a good domain for studying collaborative human/agent decision-making as they provide complexity, but also concrete right and wrong answers that can help illuminate how humans and machines should collaborate in other situations.

4. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy: Why users compromise computer security mechanisms and how to take remedial measures. *Communications of the ACM*, 42(12):40–46, December 1999.

[2] S. Afroz and R. Greenstadt. Phishzoo: An automated web phishing detection approach based on profiling and fuzzy matching. Technical Report DU-CS-09-03, Drexel University, 2009.

[3] M. Brennan and R. Greenstadt. Practical attacks on authorship recognition techniques. In *Innovative Applications of Artificial Intelligence*, 2009.

[4] Y. Cao, W. Han, and Y. Le. Anti-phishing based on automated individual white-list. In *DIM '08: Proceedings of the 4th ACM workshop on Digital identity management*, pages 51–60, New York, NY, USA, 2008. ACM.

[5] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell. Client-side defense against web-based identity theft. In *11th Annual Network and Distributed System Security Symposium (NDSS '04)*, February 2004.

[6] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *CHI '06: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2006.

[7] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *CHI '06: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2006.

[8] W. K. Edwards, E. Shehan, and J. Stoll. Security automation considered harmful? In *New Security Paradigms Workshop (NSPW)*, 2007.

[9] S. Egelman, L. Cranor, and J. Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *CHI*, 2008.

[10] M. Jakobsson et al. What instills trust? a qualitative study of phishing. In *Proceeding of first Int'l Workshop on Usable Security, Springer-Verlag*, 2007.

[11] R. Greenstadt and J. Beal. Cognitive security for personal devices. In *First ACM Workshop on AISec (AISec'08), ACM CCS 2008 Conference*, 2008.

[12] Anti-Phishing Working Group. Global phishing survey: Domain name use and trends in 1h2008. http://www.antiphishing.org/reports/APWG_GlobalPhishingSurvey1H2008.pdf, 2008.

[13] A. Herzberg and A. Gbara. Trustbar: Protecting (even naive) web users from spoofing and phishing attacks. *Cryptology ePrint Archive*, (155), 2004.

[14] E. Horvitz. Principles of mixed-initiative user interfaces. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 159–166, 1999.

[15] W. Inc. Waterken yurl trust management for humans. http://www.waterken.com/dev/YURL/Name/.

[16] G. Kacmarcik and M. Gamon. Obfuscating document stylometry to preserve author anonymity. In *COLING/ACL on Main conference poster sessions*, July 2006.

[17] P. Kelley, P. H. Drielsma, N. Sadeh, and L. Cranor. User-controllable learning of security and privacy policies. In *First ACM Workshop on AISec (AISec'08), ACM CCS 2008 Conference*, 2008.

[18] E. Michelakis, I. Androutsopoulous, G. Paliuras, and G. Sakkis. Filtron: A learning-based anti-spam filter. In *1st Conference on Email and Anti-Spam*, 2004.

[19] S. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *IEEE Symposium on Security and Privacy*, May 2007.

[20] B. Schneier. The psychology of security. http://www.schneier.com/essay-155.html, 2008.

[21] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *Eurocrypt*, 2003.

[22] A. Whitten and J. Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *Usenix Security Symposium*, 1999.