Captchat: A Messaging Tool to Frustrate Ubiquitous Surveillance

Paul Dunphy
Culture Lab
Newcastle University
paul.dunphy@newcastle.ac.uk

Patrick Olivier
Culture Lab
Newcastle University
first.lastname@newcastle.ac.uk

Johannes Schöning
Expertise Centre for Digital Media
Hasselt University - tUL - iMinds
Johannes.schoening@uhasselt.be

James Nicholson
PaCT Lab
Northumbria University
james.nicholson@northumbria.ac.uk

CHI'15 Extended Abstracts, April 18 - 23, 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3146-3/15/04…$15.00
http://dx.doi.org/10.1145/2702613.2732515

Abstract

There is currently a widespread uncertainty regarding the ability of citizens to control privacy online in the face of ubiquitous surveillance. This is a huge and complex societal problem. Despite the multi-faceted nature of the problem, we propose that HCI researchers can still make a positive contribution in this space through the design of technologies that support citizens to engage with issues of surveillance. In this paper we describe the design of a messaging application called Captchat. Captchat enables people to send everyday messages embedded into images, with the added ability to apply visual distortions to the message to resemble an online CAPTCHA. We propose the chief benefit would be that Captchat messages (with potentially "one-time" distortions) can increase the difficulty for algorithms to index private messages and necessitate the involvement of much more costly human labor in the surveillance process. We developed a prototype and conducted a user study; the results suggest that people were likely to create Captchat messages that were difficult to index for an OCR package but still easy to understand by humans, even without explicit instructions to interact 'securely' with the application. While more work is still required to understand the limitations of Captchat, we hope it can open discussion on how HCI researchers can respond to the challenges faced from ubiquitous surveillance.

Author Keywords
Ubiquitous Surveillance, Mobile Messaging, Privacy, Captchat

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Ubiquitous Surveillance is a term that is increasingly used to describe the reality in technology-advanced societies, where personal digital information is collected, stored, searched, correlated and analysed. The data that is particularly targeted with these methods is usually generated by everyday services e.g. location-based services, email, messaging. Such activities are periodically in the spotlight due to media reports detailing the surveillance capabilities of nation states and large corporations. In this new eco-system of surveillance, the chief threat to privacy comes through the application of data mining, performed by those to whom we disclose our information to, directly and indirectly. However, the sheer invisibility and highly technical nature of this kind of surveillance makes it nearly impossible to engage people with societal agendas that argue for greater respect for citizen privacy.

Personal messaging (e.g. email, instant messaging) is an application area that raises classic privacy concerns. Mechanisms to protect security or privacy online in this area have historically been too technically sophisticated to be scrutable by users who, due to their lack of expertise, are essentially asked take a leap of faith that encryption is doing something useful. So while personal end-to-end encryption is (and should be) considered a gold standard for confidentiality online, early studies showed that this was incomprehensible to a group of lay users [7]. In some cases, researchers have proposed that rejection of security or privacy features on these terms can be considered entirely rational [4]. Added to this is the knowledge that security and privacy are rarely primary concerns for users; these two points combined suggest that new technologies with cryptographic features heavily foregrounded appear unlikely to see wide personal adoption (beyond activists), despite the myriad of free technologies that exist in this space.

The case of the widely popular *ephemeral app* [2] Snapchat shows how privacy can be packaged differently: the application provides concrete and enjoyable added value in the messaging interactions of the user, and also a scrutable privacy experience. In Snapchat, users send pictures or videos to a recipient with the added ability to specify a time duration after which that message is deleted from the phone of the recipient and the server itself (from 1-10 seconds, enforced by the app). While the exact workings of this infrastructure are often debated, this app does succeed in providing a tangible experience [3] of privacy, in the form of a salient understanding of how privacy is enforced, controlled and ultimately, compromised.

The increasing interest in the design of ephemeral apps [2] suggests that users have an appetite to control and experiment with privacy in new ways that add value to existing messaging interactions and are useful in everyday life. We propose that the application presented later in this paper provides one example of a concept that could actively engage users in privacy protecting behaviors, and be fun, without the app being pitched as a security or privacy 'tool'. We also propose that the CHI community should actively design and develop such applications to support people to participate, even in small ways, in the debate around privacy in society. The goal must be to explore how the contemporary experience of privacy must look and feel, and to aim to bridge the gap with the technical reality.

Captchat

The contribution of this paper is the design and evaluation of Captchat: a messaging tool that leverages users' extensive everyday experience of CAPTCHA [1] as a technology to resist *bots,* in the design of a messaging platform. Using Captchat, users can create and distribute personal messages that are expensive for algorithms to perform text extraction upon, yet are readable by another human. In this section we provide more detail regarding the concept of the application, but also the implementation details of our current prototype that was used to drive our research in a user study.

Captchat Benefits

Captchat enables users to create everyday messages and apply classic visual distortions prominent in CAPTCHA design and send those messages purely in image format to their friends. This creates the need for prospective data miners to begin pixel level analysis to segment the textual characters and understand the message (see Figure 1). The system presents two main benefits that overcome traditional limitations of online CAPTCHA:

(i) Recipients are not required to understand every character in the message as with online CAPTCHA, but must simply be able to understand the

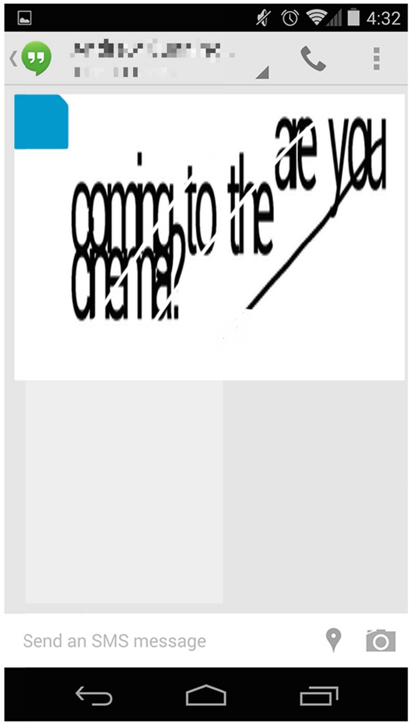

Figure 1: The Captchat application in use. A message has been received where the text can be interpreted as "Are you coming to the cinema?". The messages are "secured" by the users applying visual distortions to the textual message. These distortions could be pre-stored / pre-designed, user-defined, different for each chat partner, or created on the fly before sending the message.

message as a whole. This leverages principles of Gestalt Psychology [5] that human brains are effective at working with incomplete information;

(ii) Machine learning attacks on the content can be thwarted as visual distortions do not need to be consistent between messages, and can even be applied unpredictably to each message. The intention is that this approach can serve to create effort for methods of data mining, in a way that can still be understood by users. In the ideal case, distortions could be "one-time".

As such, Captchat aims to re-appropriate the underlying principles of CAPTCHA for message creation and distribution. This resulting message is not a CAPTCHA, nor is it steganography (the practice of hidden writing) as we do not hide the fact there is a message within the image. The message represents the result of a function that is easy for humans to apply and invert (understand), but difficult and error-prone to decode by algorithm. Messages are not stored on any Captchat server.

Captchat prototype

To instantiate the Captchat idea, we developed a prototype on the Google Android platform (version 4.4 Kitkat). The initial set of visual distortion features we included were guided by research conducted by Yan et al. [8], that discusses how particular visual distortions can effect usability and security. The features we included in this Captchat prototype are the following:

(i) Change the spacing between characters and words;
(ii) the ability to scribble over the message;
(iii) text stretching/compression
(iv) change the colour of the background or text;
(v) change the positioning of the text;
(vi) displace the ordering of the words.

The distortion set here does not exhaust the space of possibilities but provides a starting point for our studies. The distortion set must provide a suitable balance between the possibility for users to have some fun with the features, but also to provide reasonable distortions to the messages.

User Study

One important research question about Captchat relates to the types of distorted message that users would create given free choice; would these messages be relatively secure (from text mining) and readable? To explore this question further we conducted a user study. Our goal was to involve users to generate a battery of messages with bespoke distortion that would enable us to observe potential usage of the application and the relationship between readability by a person and an typical OCR package.

Participants

We recruited 18 participants to take part in the study (age: μ=25 years old). The only exclusion criteria involved uncorrectable vision deficiencies including colour blindness or other severe sight impairment. All participants were experienced using mobile devices and mobile messaging but none had significant knowledge about online CAPTCHAs or their associated security.

Method

We conducted a lab study using a within-subjects design. Each participant was firstly provided with a Nexus 5 mobile device with the Captchat application

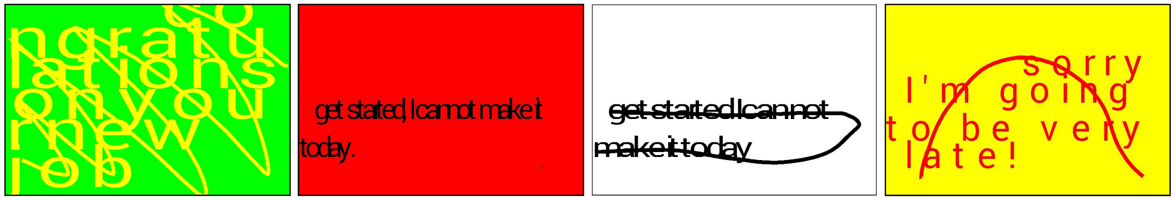

Figure 2: Examples of Capchat messages created in our user study. All messages were successfully transcribed by in the human readability study, but were not recognized by the OCR software.

pre-installed. A short explanation about the goals of the application was provided, although we did not tell participants that this was a security-related messaging application; we described the application as a messaging application that allows people to customize the visual appearance of messages, based upon visual distortions. Then participants were given time to experiment with the functionality of the system and ask any questions.

Participants were then asked to type five different pre-determined sentences (controlled for length; order balanced per participant) and distort each sentence based upon an instruction drawn randomly from a set of five (each instruction given once only): a free transformation (control), a beautiful design, a minimal design, a design difficult for a computer to read (secure), and a fun design. Participants were always asked to send the control message first, but the order of the other four conditions was balanced per participant. All participants were told that the resulting messages should be readable by a potential recipient, whilst also satisfying the given instruction.

These instructions were chosen to determine which set of message designs might lead to the best overall messages – in terms of resistance to OCR yet still readable by a human. Participants were able to use any of the design features of the application to meet these criteria.

Following the study we evaluated the messages in two ways (i) we placed the messages on the Amazon Mechanical Turk to test the readability of the message; (ii) we imported the messages into Abbyy FineReader OCR software to gain a rudimentary measure of the ease with which the text could be extracted from the message

Results

In the study we collected 90 Captchat messages, and firstly we qualitatively analysed the messages according to the visual characteristics we found across each type of message. We found that messages designed to be minimal or beautiful tended to make use of high contrasting colours e.g. black and white, centred the text, and tended to make poor usage of the features that allowed manipulation of the spacing or the scribble function.

Control	Beautiful	Minimal	Secure	Fun
44% (61%)	39% (44%)	28% (39%)	6% (11%)	0% (6%)

Table 1: Percentage of messages in each condition where text was successfully extracted by the OCR software. Brackets indicate close partial matches.

Condition	Mean	σ	95% CI
Control	1.46	0.9	(1.3, 1.6)
Minimal	1.55	1.1	(1.3, 1.8)
Beautiful	1.23	0.6	(1.1,1.4)
Fun	2.08	1.1	(1.9, 2.3)
Secure	3.12	1.6	(2.8,3.4)

Table 2: Perceived readability of Captchat messages according to the conditions under which they were created by MTurkers (higher = more difficult).

Fun designs appeared visually similar to the condition that requested secure designs, these made high usage of features that created very tight spacing effects and heavy use of the scribble functionality. Messages created in the control condition had characteristics of both groups but appeared most similar to the minimal and beautiful design groups; and those drawings tended to make little use of the colour palette.

OCR Readability

We processed each message using the Abbyy FineReader OCR software; all results are summarized in Table 1. A Pearson's chi-square test indicated that the messages sent with fun transformations had noticeably fewer correct guesses by the OCR software than the rest of the conditions, $\chi2(4)=15.776$, $p=.003$. In fact, the software was unable to extract the complete text from a single 'fun' message. On the other hand, the control messages were the easiest distortions to extract, with nearly 45% guessing rate. Although, when considering close partial matches the efficacy of the OCR package increased in all cases (see Table 1). A close partial match was defined as a case where the OCR package failed to identify *only* a single character from the whole message (e.g. 'vulnerable' as 'ulnerable').

Human Readability

To evaluate the readability of the messages that were created in our lab study, we uploaded each message onto Amazon Mechanical Turk (MTurk) for a transcription task. For each message that was placed online we asked the participant to do two things: (i) transcribe the text within the image; (ii) rate on a 5 point Likert scale the difficulty of the task of transcription. We requested that each different message would be seen by 5 people on MTurk. We found that 100% of all of the messages we collected could be successfully read and transcribed by workers on MTurk. However, a closer look at subjective ratings of transcription effort reveal that some messages were more difficult to transcribe than others (Table 2). The messages created in the fun condition were rated as being similarly difficult to read as the secure messages. In a Wilcoxon Rank Sum Test there was a significant difference in the subjective readability of the messages created in the secure condition when compared to the control condition ($Z=-7.2215$, $p<0.05$). Also there was a significant difference in the perceived readability of the messages created in the fun condition and the control group (Z= -4.3575, $p<0.05$). This suggests that messages created under those two conditions were significantly more difficult to read for a human than those created in the control condition.

Discussion

In the following sections we consider the implications of our user study and the requirements for a Captchat application that is used in everyday life.

Aligning Privacy with Fun

We discovered that the Captchat messages created by participants instructed to apply a 'fun' distortion were more resistant to being decoded in our OCR experiment than those created to be 'secure'. In addition, when testing the readability of those messages with humans on Amazon MTurk, we discovered that all of the messages could be transcribed verbatim by the online participants; this indicates no reduction in usability.

The messages created in our control group were the most vulnerable to being decoded by OCR, which highlights that the natural inclination of users is not to create secure visual CAPTCHAs. However, the other instructions that we issued to participants shed light on the relationship between the guidance the application might need to provide, and the kind of messages that might result. Such insights create interesting implications for the use of Captchat, which is unlikely to be successful in practice if positioned as a privacy or security preserving 'tool'.

Captchat in Practice

Our user study took place in a lab and focused upon understanding the features of Captchat that users might naturally use when creating messages. We observed it is unlikely that users could be expected to apply careful visual distortions to every message that they send. Research has suggested that the average user of a messaging application sends around 1000 messages per month [1]. This means that new interaction strategies must be developed to make Captchat not require distortions to be configured from scratch for every message that is sent.

A number of strategies to achieve this are possible: (i) users could define a distortion template, within which, some parameters are randomized upon sending each message; (ii) the user could be periodically requested to change the distortion template; (iii) the distortion template could naturally change over time. Further studies in the field would be necessary to understand the appropriate balance between maintaining engagement with the distortion features of Captchat (necessary to provide a peripheral awareness of the goals of the application) and enabling users to send messages quickly.

Study Limitations

The study was conducted in a lab environment where people were asked to send messages that were not privacy sensitive. A stronger impression of real-world usage would have been obtained in a field study context, which is our future work. In our study we did not conduct aggressive and bespoke character segmentation attacks on our messages, but this is a topic for future work, along with the expansion of the distortion features in the prototype.

Conclusion and Future Work

Ubiquitous surveillance is a feature of modern society that affects everyone. In this paper we introduced Captchat; a messaging application that leverages the experience people have with online CAPTCHAs as inhibitors of *bots*, in the design of a messaging application. Captchat aims to provide an engaging way to exchange messages between friends, and to support fun interactions that can – as a fortuitous side effect -- frustrate the text mining of personal messages.

In a user study we found that the visual distortion features that we provided were resistant to OCR to varying degrees, but in particular that Captchat messages created under the instructions of 'fun' resulted in similar resistance to OCR as those created under the condition of 'security'. This is an important result, as applications such as Captchat are unlikely to be widely adopted with security and privacy foregrounded as user benefits. We also uncovered that while the more elaborate visual messages were rated as being more difficult to read for humans, MTurkers were still able to transcribe the messages perfectly.

Captchat is not a panacea. However, at a time when being in control of privacy is prohibitively difficult, we propose that our research can stimulate new ideas for how the HCI community can respond to the challenges raised by ubiquitous surveillance; by designing technology to support citizens to better understand, interfere with, and show dissent towards privacy intrusive practices.

Acknowledgments

This work was funded by Research Councils UK Digital Economy theme Social Inclusion through the Digital Economy Research Hub (EP/G066019/1).

References

[1] Böhmer, M., Hecht, B., Schöning, J., Krüger, A., & Bauer, G. (2011). Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In Proc. of MobileHCI. 47-56, ACM.

[2] Döring, T., Sylvester, A., & Schmidt, A. (2013, February). A design space for ephemeral user interfaces. In Proc. of TEI, 75-82, ACM.

[3] Dunphy, P., Vines, J., Coles-Kemp, L., et al. Understanding the Experience-centeredness of Security and Privacy Technologies. In Proc. of the New Security Paradigms Workshop (NSPW), (2014).

[4] Herley, C. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In Proc. of the Workshop on New Security Paradigms Workshop, ACM (2009), 133–144.

[5] Koffka, K. Principles of Gestalt psychology. Routledge, 2013.

[6] Von Ahn, L., Blum, M., Hopper, N.J., and Langford, J. CAPTCHA: Using hard AI problems for security. In Proc. of EUROCRYPT 2003. Springer, 2003, 294–311.

[7] Whitten, A. and Tygar, J.D. Why Johnny can't encrypt: a usability evaluation of PGP 5.0. In Proc. of USENIX Security Symposium - USENIX Association (1999)

[8] Yan, J. and El Ahmad, A.S. Usability of CAPTCHAs or Usability Issues in CAPTCHA Design. In Proc. of the *Symposium on Usable Privacy and Security*, ACM (2008), 44–52.