Metaphors considered harmful? An exploratory study of the effectiveness of functional metaphors for end-to-end encryption

Albesë Demjaha, Jonathan Spring, Ingolf Becker, Simon Parkin and M. Angela Sasse
University College London
{albese.demjaha.16, jonathan.spring.15, i.becker, s.parkin, a.sasse}@ucl.ac.uk

***Abstract*—Background: Research has shown that users do not use encryption and fail to understand the security properties which encryption provides. We hypothesise that one contributing factor to failed user understanding is poor explanations of security properties, as the technical descriptions used to explain encryption focus on structural mental models.**
Purpose: We methodically generate metaphors for end-to-end (E2E) encryption that cue functional models and develop and test the metaphors' effect on users' understanding of E2E-encryption.
Data: Transcripts of 98 interviews with users of various E2E-encrypted messaging apps and 211 survey responses.
Method: First, we code the user interviews and extract promising explanations. These user-provided explanations inform the creation of metaphors using a framework for generating metaphors adapted from literature. The generated metaphors and existing industry descriptions of E2E-encryption are analytically evaluated. Finally, we design and conduct a survey to test whether exposing users to these descriptions improves their understanding of the functionality provided by E2E-encrypted messaging apps.
Results: While the analytical evaluation showed promising results, none of the descriptions tested in the survey improve understanding; descriptions frequently cue users in a way that undoes their previously correct understanding. Metaphors developed from user language are better than existing industry descriptions, in that ours cause less harm.
Conclusion: Creating explanatory metaphors for encryption technologies is hard. Short statements that attempt to cue mental models do not improve participants' understanding. Better solutions should build on our methodology to test a variety of potential metaphors, to understand both the improvement and harm that metaphors may elicit.

I. Introduction

The purpose of cryptography has expanded from benefiting primarily the military to securing systems for the general public [1]. For example, widely used messaging applications such as WhatsApp and Telegram have embraced the use of end-to-end (E2E) encryption. This recent flourishing of technology gives non-expert users free access to encrypted person-to-person communication. That general users can take advantage of encryption is however in question - in 1999, Whitten and Tygar [2] claimed that cryptographic details confuse users. Concerns persist about the usability of E2E-encryption, and continue to generate a considerable body of research [3]–[7].

Workshop on Usable Security (USEC) 2018
18 February 2018, San Diego, CA, USA
ISBN 1-1891562-53-3
http://dx.doi.org/10.14722/usec.2018.23015
www.ndss-symposium.org

Saltzer and Schroeder [8] have long encouraged designers to bridge the gap between a user's mental image of a protection system and the system's specification language. However, few attempts have been made to address this gap. Application properties are communicated to users in technical jargon, potentially obstructing their comprehension of security features (including *encryption*, which itself is a technical cryptographic term). The focus of communication efforts appears to be to teach users how the system works, with the tacit assumption that improved understanding will allow users to complete their tasks. In "Analogy Considered Harmful" [9], Halasz and Moran argue that analogies by themselves are insufficient for teaching people about computing systems. We expand their ideas to metaphors and provide empirical evidence.

Beyond cryptographic tools, systems can be explained to users through accustomed metaphors [10], which can be extended into design models. DiSessa [11] distinguishes between a 'structural' and a 'functional' mental model. A structural model provides a detailed understanding of the system, whereas a functional model provides certain properties of the system which are necessary to complete a real-life task.[1] After Whitten and Tygar [2] concluded in their *'Johnny'* paper that the 'key' metaphor was misleading, Whitten [13] responded by visually transforming the 'key' metaphor for a secure e-mailing tool. This was effectively a revised structural model intended to cue users' understanding of cryptographic functions. Whitten does not question the appropriateness of the 'key' metaphor itself, nor whether the strategy of cuing structural models will lead to better user outcomes.

There is a dearth of studies that specifically address the issue of inadequate terminology and metaphors to describe cryptographic systems such as E2E-encryption. Furthermore, no attempts have been made to generate and test metaphors in a comparable, repeatable manner. Here we break this impasse and adapt HCI methodology to rigorously generate metaphors for E2E-encryption. We aim to generate metaphors that cue functional mental models in users [11]. Rather than requiring users to explicitly learn all the relevant security tasks, we aim to cue users' already-existing models. Metaphors leverage these existing models to approximate a working mental model for performing a specific task.

[1] A functional model is similar to a task-action mapping model – defined by Young [12] as an internalised representation of the system to the real-world task which users have to perform.

The research aims of this paper are: (1) *to rigorously generate metaphors that cue functional mental models*, and (2) *to test the effectiveness of these metaphors with users*.

We present a modified application of the framework for re-engineering metaphors by Alty et al. [14]. Our first contribution is our proposed methodology which serves as a guide for the generation of explanatory metaphors. It suggests a combination of analytic evaluation and empirical testing which provides useful evidence on the efficacy of the metaphors. Furthermore, we generate five new metaphors for E2E-encryption, further refined to a set of three. Thirdly, we test these three metaphors and two currently used explanatory metaphors through a survey of users of communication apps containing E2E-encryption capabilities.

The results of the survey show evidence that currently used metaphors harm users' understanding of secure messaging app functionality such as E2E-encryption. Moreover, our results show that metaphors generated from user language do not cause such harm to user understanding. However, our results are negative in that none of our new metaphors actually cue a 'correct' mental model in users. Extended metaphors may be needed to more effectively cue functional mental models of E2E-encryption. Our methodology can be used to better understand both the improvement and harm that metaphors may cause to users' understanding of E2E-encryption.

The rest of the paper is organised as follows: The upcoming section summarises a corpus of metaphor research in security and identifies the research gap which motivates the paper. We then describe our four-step methodology in Section IV, followed by a summary of our results in Section V. Our findings are further examined in Section VI, before we conclude with recommendations and make suggestions for future work.

II. Related Work

We present related work in two areas: usable E2E-encryption cryptography and understanding of user mental models in security.

A. Usable E2E-encryption Cryptography

The effectiveness of E2E-encryption solutions for non-expert users continues to be studied in the research community [3]–[7]. In *'Why Johnny Can't Encrypt'* [2], Whitten and Tygar found that PGP 5.0's graphical user interface (GUI) was not as usable as advertised. Most participants were unable to correctly send an encrypted e-mail. The authors concluded that PGP 5.0 had an insufficiently usable GUI – including a signature and pen metaphor – that led to dangerous errors. Whitten sought to correct this in studying Lime, a proposed secure e-mailing tool [13]. Her argument was that (i) effective metaphors for security are lacking, and (ii) metaphors need to map to critical process information if they are to support users. Regardless, Whitten stayed focused on structural components of the mental model, such as keys and locks in the right places.

There is no consensus on what exactly should be improved, although technical solutions have not been sufficient. Garfinkel and Miller [15] conclude PGP failed due to inadequate key certification model, not its GUI. However, even after automating these key-based tasks, they found PGP security problems remained. Subsequently, Sheng et al. [16] continued to identify key certification and UI issues in PGP. Fahl et al. [3] find users need to feel security features 'do something'. However, this is a delicate balance, as Ruoti et al. [17] find users are confused when security details are too transparent. Ruoti et al. [18, p. 4] advocated 'approachable descriptions of public-key cryptography'. Most modern E2E-encryption messaging apps instead hide such structural details from the user. This design decision perhaps accepts a failure – that indoctrinating users with structural details did not work well.

But all is not well with modern apps. There are two modes of E2E-encryption underlying IM applications *opportunistic* and *authenticated* E2E-encryption [5]. Users are often unaware of the difference. In response, Abu-Salma et al. suggest that *'security properties and threats should be framed in terms that users can understand'* [7, p. 15]. When terminology is inappropriate, the incidence of insecure behaviour increases. Next, we explore explicit attempts at providing for users an approachable description of encryption.

B. Metaphors and Mental Models in Security

There are few attempts to find appropriate terminology to engage users with computer systems. Clark and Sasse [10] applied conceptual design to create a user interface for the Session Directory Tool (sdr). Users' existing knowledge and context of use was elicited, identifying metaphors which were then adapted into a design model. The other major work is Whitten [13], in which she develops a technique for tailoring visual security metaphors. Whitten visually tailors the traditionally used public-key cryptography metaphors and incorporates them into a new secure e-mailing tool. She tests her metaphors empirically by providing users with a detailed description of Lime accompanied by the visual metaphors. Results show that users do not particularly perform better when using Lime rather than PGP, despite the tailored metaphors.

Following Whitten, other authors have proposed visual metaphors for secure e-mailing tools. Roth et al. [19] suggested mail envelopes and postcards. Tong et al. [20] proposed metaphors arguably couched in the jargon of domain experts, such as key, lock, seal and imprint. Lausch et al. [21] revisited postcard and mail envelope metaphors, but extended them to include a torn letter to signify a corrupted e-mail. Although the authors presented promising survey results, their participants were 'privacy aware' and 'tech-savvy.' There is little indication such metaphors similarly impact non-experts, ignorant of encryption technology operation.

Unlike Whitten, Clark and Sasse involved users in generating metaphors [10]. They found that a metaphor grounded in users' own experiences could leverage simple models of understanding to support users who lack the knowledge of how underlying systems work. Users' tool facility then approaches the competence of an experienced user with domain knowledge.

Users' mental models of a system are usually simpler than the real-world system [22], [23]. Johnson-Laird's theory of mental models originates in cognitive psychology, and is used in human-computer interaction (HCI) [10], [23] and recently, security. When generating metaphors, one must consider users'

previous knowledge of the intended system. Renaud et al. advise security researchers to *'nurture and foster comprehensive and complete mental models of E2E to ensure that users want to encrypt, know how to encrypt and, most importantly, do encrypt'* [24, p. 17].

Users tend not to have an accurate model of security threats or security tools. Wash [25] finds home users often choose to ignore advice from security experts about threats. User understanding of threats such as bot-nets bears little resemblance to the technical reality. Abu-Salma et al. [7] find users have poor mental models of secure communications, how encryption works, and E2E-encryption.

There is evidence that media channels are finding information security terminology, and cryptographic terminology in particular, a barrier to communicating to their audiences. The 'Planet Money' podcast [26] for example discusses *'some Russian hackers using an electronic device to mess with the outcomes of a slot machine'*. 'The Ceremony' [27] describes use of Zcash crypto-currency as a ceremony containing math and wizardry. These efforts to make security accessible are well-intentioned. However, the media's explanatory goals are different from those of practitioners. One is not a substitute for another. This may explain why the media is potentially 'dumbing down' the technology rather than providing their audience with the skills to move ahead in a more secure way.

'The Analogies Project' [28] blog also makes an effort to explain security concepts to users through analogies such as *Infosec is Like Sun Protection* [29]. However, the majority of these analogies do not serve the purpose of an analogy because they do not cue understanding of security concepts in a self-contained manner. Practitioners should aim to provide adequate explanatory metaphors.

Twenty years ago, Clark and Sasse [10] argued for adequate explanatory metaphors as an aim. Specifically, to enable novice users of encrypted communications to appear skillful in their use of complex tools. The practitioner community has yet to find the right metaphors that enable such skilled use. We present a methodology for improving the community's pursuit of this difficult task. We also take a different direction from past work that focuses on instilling detailed structural models; instead we attempt to cue functional models.

III. Background

This section provides a brief background of Alty et al.'s [14] metaphor evaluation framework and the specific parts of the framework that we use in this paper.

A. The Framework

The framework that we apply consists of an extensive process for generating and evaluating metaphors at the user interface (UI) [14]. Due to contextual differences, we make some alterations to the framework and apply a concise version. For example, in Alty et al.'s scenario, some of the steps involve the integration of the metaphor into the user interface of a system. Although our study has user-metaphor interaction, it lacks user-metaphor interaction through the UI of a secure messaging app. Thus, the steps that we apply from the framework are: identification of system functionality; generation and description of potential metaphors; analysis of metaphor-system pairings, and; evaluation. We identify the system functionality in IV, elaborate the generation and description of the metaphors in IV-A and IV-B, and show their evaluation in IV-D.

B. Analysis of metaphor-system pairings

The matrix used to analyse metaphor-system pairings was originally presented by Anderson et al. [30] within the above mentioned framework and it is based on their vast experience of creating metaphors for telecommunications systems. Corresponding to the aim of the framework, it serves as a guide to software designers developing metaphors to describe computer systems. We have recreated the contents of the original matrix in Table I below. The authors borrow terminology from psycholinguistic literature – they use the word *vehicle* to represent the metaphor and the word *system* to refer to the computer system being analysed [30].

TABLE I. Interaction between set of vehicle features and set of system functionality — reproduced from [30]

	V+	V-
S+	Those features provided by the system and supported by the vehicle (S+V+)	Those features provided by the system but not supported by the vehicle (S+V-)
S-	Features implied by the vehicle but not supported by the system (S-V+)	Features not implied by the vehicle and not supported by the system (S-V-)

Brostoff et al. [31] refactor the original matrix shown above, and apply it within the field of security. They refer to it as *The Metaphor Evaluation Matrix* – the word 'vehicle' is substituted with the word 'metaphor'. We adopt Brostoff et al.'s version of the matrix and change it accordingly (see IV-C).

IV. Methodology

Our methodology consists of four parts:

1) Using interview transcripts to explore potential user explanations that can be used as competent metaphors (Section IV-A)
2) Applying a modified version of Alty et al.'s [14] framework to generate new metaphors (Section IV-B)
3) Analytically evaluating the new metaphors through the same framework in order to select promising ones (Section IV-C)
4) Designing and conducting a survey for testing whether the selected new metaphors cue better understanding of E2E-encryption in the context of secure messaging apps (Section IV-D)

Before we describe these four parts in detail, we must make a choice of definitions. Alty et al. [14] use the term 'system' to refer to the computer system to be described through a metaphor. In this paper, the system is *E2E-encryption*. Our baseline definition for E2E-encryption is taken from RFC 4949, the Internet Security Glossary:

> "Continuous protection of data that flows between two points in a network, effected by encrypting data

when it leaves its source, keeping it encrypted while it passes through any intermediate computers (such as routers), and decrypting it only when it arrives at the intended final destination" [32, p. 121].

We also must define which criteria of the system are most salient and important. We attempt to strike a balance between a definition that is detailed enough to cover the main purposes of E2E-encryption, and also sufficiently simple for relatively quick analytic assessment by a prospective researcher. For example, in order to implement E2E-encryption, the developer must make decisions about key management, such as forward secrecy, and traffic analysis resistance, such as by traffic padding. Assessment regimes such as NIST's key management recommendations [33] and security engineering recommendations [34] provide a great amount of detail. However, we cannot evaluate metaphors on the details of many-hundred pages of recommendations. Our goal was to identify properties that allow users to reach their security goals when interacting with a secure messaging app. We focus on E2E-encryption itself, not resisting traffic analysis or the extant threat ecosystem. Measuring and effectively cuing user understanding in these three areas is material for future work. We identify the core evaluation criteria as:

Safe Implementation: The contents of any message or phone-call should not be retrievable without knowledge of the key

Confidentiality: Transmitted data should appear unintelligible to anyone outside the conversation

Coverage: Data should always be encrypted from its source to its destination

Authentication: The contents of any message or phone-call should not be retrievable by an adversary in the middle, masquerading as the intended recipient

Key Management: The key should merely be shared by the two trusted endpoints

A. Interviews with users of E2E-encrypted messaging apps

We conduct secondary thematic coding of interview transcripts originally analysed in three separate user studies, for the purpose of extracting candidate metaphor topics. Study 1 [7] explores user behaviour, perception, and understanding of secure communications (58 transcripts). Study 2 [6] explores user adoption and understanding of Telegram and evaluates the tool's user interface (20 transcripts). Study 3 [35] captures user perceptions of WhatsApp's security properties in a mock-up containing modifications of certain features (20 transcripts).

We extract user descriptions of encryption and E2E-encryption which are independently coded by three researchers for completeness or incompleteness as well as technical or non-technical accuracy. We assess the inter-coder agreement for the coding of the encryption and E2E-encryption quotations separately. Krippendorff's alpha [36] coefficient is 0.612 for encryption, and 0.723 for E2E-encryption. Generally, a coefficient above 0.66 is considered sufficiently reliable. However, since for the set of encryption quotations our coefficient is below 0.66, we look at pairwise numbers which highlight a deficiency of one coder. Therefore, we finalise the codebook by taking the majority agreement and including codes on which at least two coders agree. Our final codebook consists of 183 valid quotations. The distribution of the codes is shown in Table II.

TABLE II. DISTRIBUTION OF INTERVIEW CODES

	Encryption	E2E-Encryption
Lack of knowledge	25	9
Incorrect technical description	10	2
Incomplete technical description	9	5
Sufficient technical description	0	2
Incorrect non-technical description	43	9
Incomplete non-technical description	43	16
Sufficient non-technical description	8	2

The quotations from our codebook further justify the need for exploring alternative explanatory metaphors that cue better understanding of E2E-encryption in users. From the quotations attempting to explain E2E-encryption, there were only four descriptions which are labelled as 'sufficient' by all three coders.

The quotations coded as sufficient technical descriptions are:

1) "Err my understanding of it from what I've been told anyway is that it is erm the messages are encrypted at both ends. So it's like if I'm sending a message it encrypts the message sends it to the other user and then they can decrypt it rather than anything being sent over plain text, so no keys are passed across plain text or anything. Erm and it's just ... it's yeah it's more difficult to crack because anyone intercepting the message needs to break it somewhere along the way there's no way to see it."
2) "It means when I send it, it will be coded in some password and when my contact receives it, it will be decoded and no one can steal the message in between I guess."

The quotations coded as sufficient non-technical descriptions are:

1) "According to my limited understanding, is a security kind of system where only you and the receiver would be able to read the messages and nobody could intercept."
2) "Whoever I've sent the message to, so here only my contact can see what I've written, but anyone else other than me or my contact can't, I think."

These four statements indicate some end-users can produce adequate explanations. Unfortunately, none seem like a clear way to convey better mental models. The technical descriptions just use the jargon from the industry adequately; we will test this approach by testing metaphors used in apps with E2E-encryption. The two non-technical descriptions each have two important features – correctly identifying the endpoints as the sender and receiver, and that no one else can intercept the message. These features will recur in the metaphors we create. However, there is not much in the statements, so there is little to concretely inspire our metaphor creation.

The three-annotator coding of the user descriptions was also used to develop an awareness of the content of the data, and to develop a useful categorisation of potential metaphors. The non-technical descriptions are re-coded by one researcher for keywords that serve for generating metaphors. Table III displays the keywords used.

TABLE III. USER KEYWORDS FOR E2E-ENCRYPTION

Keyword	Count
CODE	22
JUMBLE	7
SECRET	5
RANDOM	5
PRIVATE	5
PASSWORD	5
LETTER (A-B-C)	5
HIDDEN	5
BOX	4
UNREADABLE	3
SCRAMBLE	3
BINARY (101101)	3
SYMBOL	2
DIFFERENT LANGUAGE	2
CHARACTER	2
UNPACK	1
PICTURE	1
INVISIBLE	1

B. Generation and description of E2E-Encryption metaphors

Alty et al. [14] propose several approaches for generating metaphors. We adopted three of these approaches: *Design Metaphors*, *Brainstorming* and *Extension*. Design Metaphors emphasises the role of users as a useful source of metaphors. Users often apply familiar metaphors from their everyday life to their language to aid their understanding of a system when undertaking a task [14]. Similarly, when some participants of the user studies describe encryption or E2E-encryption, they rely on familiar terminology to aid their understanding of the system. Thus, we use the coded keywords when generating some of the new metaphors through the framework.

Brainstorming suggests mapping real-life functionalities to the functionalities of the technical system in order to identify potential metaphors. Finally, Extension promotes the idea of recycling metaphors that are currently used by similar computer systems and extending them in a manner that appropriates the metaphor to the new system.

We introduce five new metaphors in total. All metaphors are produced through one of the three above mentioned approaches. The metaphors *Special Language* and *Treasure Hunt* are a product of the Design Metaphor approach, the metaphors *Colours* and *Banknote* are a product of the Brainstorming approach, whereas *Owl* is a product of the Extension approach.

Special Language: *Messages and calls with this person will be translated to a special language for which only the two of you know the dictionary.* Participants used the word 'language' to refer to encryption. Thus, this metaphor reflects participants' statements:

- "...information is sent over the internet, but it's encrypted, so sensitive information is being sent in a different language, it's not being sent as is."
- "...you have written something and encryption means that, yes yes it's like turning the language into something else for reading the other device..."

Treasure Hunt: *Messages and calls exchanged with this person are like a treasure hidden in a place to which only the two of you know the map.* This metaphor is also generated based on participant language referring to encryption as *a message in a box* or *a message that can be unlocked*:

- "...I think it's in a box or something according to my understanding..."
- "So what I think it's more like hiding your what's inside the way that you know something is delivered but you don't know what's inside of the box."
- '...You need a password to unlock the message?..."

Colours: *Messages and calls you exchange with this person are like colours. Before sending them, you mix them with another colour, known only by you two. Nobody else can retrieve them unless they know the secret colour.* We generate this metaphor with the purpose of eliciting participants' mental models through every day talk [37]. The concept of colours is relatively simple and familiar to a general public of most ages.

Banknote: *Messages and phone calls shared with this person are matched like a ripped banknote, each piece being owned by one of the two people, therefore in order to access the message both pieces are needed.* The idea for this metaphor emerges due to the uniqueness of the 'ripped' pattern created when tearing a banknote. If two people rip a banknote and each keep the corresponding piece, it is almost certain that no other half will match.

Owl: *Messages and calls with this person will be delivered by your owl which will not share them with anyone else but the two of you.* This metaphor is somewhat based on the traditional 'mail' metaphor used within cryptography.[2]

C. Analytic evaluation of metaphors

Alty et al. [14] suggest a two-fold evaluation of metaphors; an analytic evaluation, and an evaluation that captures the metaphors' relationship with the underlying system functionalities (IV-D). The analytic evaluation is done by determining the intersection between *the system* set *S* and *the metaphor* set *M*. More specifically, it is based on: features that intersect between the two sets *M+S+*, features of the system that the metaphor does not cover *M-S+*, features of the metaphor that do not correspond to any system functionalities *M+S-*, and features that do not belong to either set *M-S-* [14]. The *M+S-* category is conceptual baggage,[3] which can lead users to incorrect assumptions about the system functionalities.

Here, set *S* contains the functionalities of E2E-encryption, whereas set *M* contains the processes of a metaphor mapped to the system functionalities. The evaluation is done through an adaptation of Brostoff et al.'s matrix [31]. Thus, *M* remains the same, whereas set *S* is referred to as set *E* which signifies the system in use, that is, *E2E-encryption*. The properties of the matrix are shown in Table IV.

Each metaphor is ranked qualitatively on the Metaphor Evaluation Matrix. Two researchers rank each separately in order to avoid favouring a particular metaphor. The elimination is based on the metaphors' coverage of system functionalities. However, set *M+E-*, i.e. 'conceptual baggage' also affects the elimination process. Even if a metaphor covers the majority of the system functionalities, a high level of conceptual baggage will likely lead to an inefficient use of the system [10]. In

[2]It is also significantly inspired by the *'Harry Potter'* universe by J.K. Rowling, which are characteristic to British culture.

[3]"Conceptual baggage can be thought of as features of a metaphor that are not utilised in a particular metaphor-system pairing" [30, p. 4].

TABLE IV. THE METAPHOR EVALUATION MATRIX

	M+	M-
E+	**Desirable** Features provided by E2E-encryption and supported by the metaphor. Leads to correct understanding and use of E2E-encryption.	**Undesirable** Features provided by E2E-encryption but not supported by the metaphor. Leads to misunderstanding and underused features of E2E-encryption.
E-	**Very undesirable** Features implied by the metaphor but not supported by E2E-encryption: Conceptual baggage Leads to user errors.	**Not important** Features not implied by the metaphor and not supported by E2E-encryption.

addition to the new metaphors, we analytically evaluate two existing descriptions used in E2E-encrypted messaging apps. We choose these descriptions based on the popularity and user adoption of the tools in which they are used.

Telegram/Viber: *Secret chats have/use end-to-end encryption.* - This metaphor shows up both on Telegram and Viber when creating secret chats.

WhatsApp: *Messages to this chat and calls are now secured with end-to-end encryption, which means WhatsApp and third parties can't read or listen to them.* - This metaphor shows up on WhatsApp conversations after tapping the initial description for 'more information'. The initial description (*"Messages to this chat and calls are now secured with end-to-end encryption"*) is almost identical to Telegram/Viber. Therefore, we decide to include the extended version to see whether it cues better understanding in users.

The metaphors that score highest on the matrix are Special Language and Colours (both equally). Treasure Hunt, Banknote and Owl cover the same number of functionalities but differ in terms of which functionalities are represented and on potential conceptual baggage. The conceptual baggage is derived intuitively and is therefore inconclusive. Following the matrix elimination, the three metaphors that we choose to evaluate further are Special Language, Colours, and Treasure Hunt. Specifically, their conceptual baggage is as follows:

- *Special Language* — users may think that the special language is picked from a set of widely spoken languages
- *Colours* — colours do not usually unmix well
- *Treasure Hunt* — messages and phone-calls appear to be hidden

In addition, we evaluate the existing metaphors for Telegram/Viber and WhatsApp. Telegram/Viber scores worse on the matrix than the disqualified new metaphors Banknote and Owl, but produces an equal amount of conceptual baggage. WhatsApp scores better than Telegram/Viber in terms of functionality coverage but has the same level of conceptual baggage. The conceptual baggage that we identify is:

- *Telegram/Viber* — (1) the word encryption is overloaded in pop culture with movie references of magic and science-fiction which may mislead users' understanding of it, (2) the wording secret chat is misleading, and (3) users may associate the security of the chat with Telegram or Viber itself rather than the functionalities of E2E-encryption.
- *WhatsApp* — (1) users may associate the word encryption with movie references which are unrealistic, and (2) users may associate the security of the chat with WhatsApp itself rather than the functionalities of E2E-encryption.

TABLE V. ANALYTIC EVALUATION OF METAPHORS

	Safe Implementation	Confidentiality	Coverage	Authentication	Key Management	Conceptual Baggage
Special Language	✓	✓	✓	✓	✓	✓
Treasure Hunt	✓	✗	✗	✓	✓	✓
Colours	✓	✓	✓	✓	✓	✓
Banknote	✗	✗	✓	✓	✓	✓✓
Owl	✓	✗	✓	✓	✗	✓✓
Telegram/Viber	✗	✗	✓	✗	✗	✓✓✓
WhatsApp	✓	✗	✓	✓	✗	✓✓

Regardless of the analytic scores of the Telegram/Viber and WhatsApp descriptions, we evaluate them further in order to empirically compare them to the new developed metaphors. Table V summarises the metaphor scores on the Metaphor Evaluation Matrix. It ranks them based on conceptual baggage and coverage of the five functionalities of E2E-encryption (Safe Implementation, Confidentiality, Coverage, Transport, Key Management) which are defined in Section IV.

D. Survey testing of metaphors

In addition to an analytic evaluation, the metaphors should be tested through interaction with users [14], [38]. In order to do this, we design a survey to test whether our new metaphors cue better understanding of E2E-encryption than the existing explanatory metaphors. Within the survey we test the following metaphors: Special Language, Colours, Treasure Hunt, Telegram/Viber, and Whatsapp.[4] We use LimeSurvey for designing the survey and the crowd-sourcing platform Prolific for recruiting participants and distributing the survey.

As our study is purely observational and does not involve any sensitive or personal identifiable data, it is considered a service evaluation by our institution's Research Ethics Committee (REC) guidelines and is exempt from REC review.

We target a population of users that reside in the United Kingdom, and have heard of at least one messaging tool that adopts E2E-encryption. The demographics that we have collected are age, level of education, adoption of messaging tools, and frequency of use.

E. Functionalities and Non-functionalities

To primarily evaluate participants' current understanding of E2E-encryption, they respond to four statements with either true or false; two of the statements are functionalities of E2E-encryption, the remaining two are not.

[4]To avoid biases in the survey we remove the words 'Telegram', 'Viber' and substitute the word 'WhatsApp' with 'application makers' in the respective industry descriptions.

Functionalities:

- Statement 1: *Only you and the recipient can read your messages* (True)
- Statement 2: *Other people can send a message pretending to be you* (False)

Non-functionalities:

- Statement 3: *Only you and the recipient can know the messages were sent* (False)
- Statement 4: *If somebody hacks your phone, they will be able to read your messages* (True)

The purpose of the chosen survey statements was to test understanding of E2E-encryption and simultaneously explore conceptual baggage. Thus, Statement 1 (True) and Statement 2 (False) were chosen to reflect functionalities of E2E-encryption. On the other hand, Statement 3 (False) and Statement 4 (True) were included as conceptual baggage because they are non-functionalities of E2E-encryption. The latter allows us to test whether the metaphors over-promise.

F. Survey Process

Participants followed the following process (a static html version of the survey can be found in the supplementary data as described in the final section of the paper):

1) Select which messaging tools they use and how frequently they use them;
2) Respond to the four statements above (true/false, in random order) in the context of using E2E-encrypted messaging applications in general;
3) Read one randomly selected metaphor and answer whether they have encountered the metaphor previously, and if yes, where;
4) Repeat the step of responding to the same four statements outlined above with true or false. The metaphor is still being shown. This evaluates whether the metaphor cues a change in the participants' understanding of E2E-encryption;
5) Ask for any comments or feedback.

G. Reliability and Validity

It is crucial to design a survey that is both reliable and valid. Reliability refers to 'the extent to which repeatedly measuring the same property produces the same result' whereas validity refers to 'the extent to which a survey question measures the property it is supposed to measure [39, p. 6]. To reduce ambiguity in the deployed survey, we conducted several rounds of revision of the questions and descriptions included in the survey, showing them to several pilot participants in each stage. Furthermore, we explore different ways of testing changes in the participants' understanding of E2E-encryption. We give one metaphor per participant to prevent the result of one metaphor impacting the result of another. In addition, by measuring a change in understanding from their initial understanding, we can test whether the results are dependent on prior understanding rather than on the metaphors.

Unintentionally introducing biases is relatively easy when designing a survey. We take a number of steps to ensure that biases are minimal:

1) Only one metaphor is allocated per participant
2) The allocation of a metaphor to a participant is random
3) The order in which each statement appears is random (both between surveys and within the same survey)

V. Results

This section presents a general overview of the survey demographics (V-A) as well as a summary of the statistical results (V-B).

A. General Overview

We collect a total of 211 valid responses from the survey. One participant is disqualified because of not using any messaging apps and 19 responses are not considered because they come from non-unique IP addresses. All five metaphors appear in the survey: Special Language (39 participants), Colours (47 participants), Treasure Hunt (48 participants), Telegram/Viber (41 participants), and WhatsApp (36 participants). From the survey participants, 57 are male and 153 female. Their ages range from 18 to 64 (average age 35). Two of our participants had no formal qualification, 36 have gone to secondary school, 73 have gone to college, 72 hold an undergraduate degree, 24 have a graduate degree, and two of our participants hold a doctorate degree.

Only 16 participants say they have seen *Telegram/Viber* before and when asked where, their answers include: Facebook (2), WhatsApp (8), BBC News (2), terms and conditions of WhatsApp (1), Facebook Messenger (1), WhatsApp group screen (1), and on a news site like BBC (1). Similarly, only 18 participants say that they have seen *WhatsApp* before and when asked where, almost all answers are associated with WhatsApp (16), in addition to: in app description or updates (1), Facebook (1), and Facebook Messenger (1). One participant says they have seen the *Treasure Hunt* metaphor online. Two participants say they have seen the *Special Language* metaphor on WhatsApp. Two participants say they have seen the *Colours* metaphor on a news story talking about encryption and on tech websites discussing secure messaging.

B. Statistical Results

Table VI compares the changes in participants' responses to the four statements due to seeing the descriptions. The first four rows indicate the difference in the number of correct responses to the statements. For example, the first 12% indicate that 12% more participants answered statement 1 correctly after seeing the Telegram/Viber description than before seeing the statement. Statistical significance is indicated by ** for $p < 0.01$ and * for $p < 0.05$, calculated by a Fisher's exact test. There are a number of interesting trends to note: in general (and especially for the *WhatsApp* description), participants' understanding of the applications functionality (Statements 1 & 2) were improved, while the non-functionality statements suffered. The *Treasure Hunt* and *Special Language* metaphors appear to be most balanced in the effect on participants understanding.

TABLE VI. DISTRIBUTION OF THE CHANGES IN THE PARTICIPANTS' RESPONSES

Statement	Telegram / Viber	WhatsApp	Treasure Hunt	Special Language	Colours
1	12%	25%	$6\%^{**}$	$8\%^{**}$	$15\%^{**}$
2	$0\%^{**}$	$11\%^{**}$	$-4\%^{**}$	$-3\%^{**}$	$0\%^{**}$
3	$-5\%^{**}$	$-11\%^{**}$	$4\%^{**}$	$13\%^{**}$	$-4\%^{**}$
4	-12%	$-25\%^{*}$	$-4\%^{**}$	$-10\%^{**}$	-11%
mean	-5%	0%	2%	8%	0%
negative mean	$-52\%^{**}$	$-53\%^{**}$	-25%	-18%	$-30\%^{*}$
changes	0.90^{**}	1.06^{**}	0.56^{**}	0.59^{**}	0.60^{**}

This is followed by the mean of the previous scores for each description. Here, a Wilcoxon signed-rank test was performed, but no description displayed statistically significant variations. While for individual statements the descriptions show positively and negatively statistically significant variations, these variations balance out upon aggregation. Hence we find no evidence that any of the descriptions give rise to an improved understanding of E2E-encryption.

However, when we penalise descriptions for causing participants to change their previously correct responses — in effect undoing the existing mental model — descriptions 1 (Telegram/Viber), 2 (WhatsApp) and (less so) metaphor 5 (Colours) appear to cause harm (row *negative mean*, Wilcoxon signed-rank test).

The bottom row indicates the mean number of responses that participants have changed after seeing the metaphor (with a maximum of 4). The mean participant has changed on average 0.73 of their responses (out of a maximum of 4 changes, statistically significantly according to a Wilcoxon signed-rank test at $p < 0.01$ for all descriptions). This confirms that participants have changed their perceptions in response to the statements, and that the results are not based on a small subset of participants.

We also perform pairwise tests for each of the statements' change scores. Here we find not a single statistically significant result, indicating that in pairwise testing no description outperforms any other (the sample size exceeds the requirements for observing a medium effect size ($f = 0.25$) in pairwise testing). This does not contradict the variations described in Table VI, as those tests were performed in-sample. So while there is evidence that some descriptions cause harm, we cannot conclude that some descriptions actually perform better than others in improving understanding of E2E-encryption.

We find a correlation between over-promising the capabilities of E2E-encryption and this decrease in performance. Essentially, the industry descriptions are more likely to make participants believe E2E-encryption does more than it actually can provide. The WhatsApp and Telegram/Viber descriptions carry more conceptual baggage than the other tested metaphors. Since baggage refers to properties of the metaphor not in the system, this is one possible explanation. A hypothesis is that the industry metaphors are more likely to influence incorrect answers post-exposure for the two questions which had a correct answer of 'no' (statements 2 and 3). The change scores for these two metaphors and statements are: 0, -2 and 4, -4 (Table VI), giving an aggregate of -2. A Fisher's exact test supports the hypothesis with $p < 0.01$ and a large effect (Cramer's $V = 0.55$).

Lastly, our analysis indicates that there is no statistically significant correlation between the participants' age, their frequency of use, and if they claim to have seen the statement previously and the descriptions' change scores, i.e. none of the additionally captured demographics and app usage statistics have any impact on the participants' responses.

VI. DISCUSSION

The discussion includes three sections. The first discusses how our results bear on our immediate research aims. The second explores the wider importance of the work and use of metaphors in explanation. The third presents limitations of our work.

A. Revisiting immediate aims

Our first aim has been to generate metaphors that cue functional mental models in users through a rigorous method. We have outlined more actionable criteria for gathering evidence about textual metaphors for explaining to end-users what E2E-encrypted messaging apps can do. Although our process can be improved and expanded in future work, our strength is multiple modes of understanding user perception of E2E-encryption. These multiple modes – interviews, analytic requirements, and survey – enable a more complex explanation of user understanding. Secondly, our survey provides evidence that the explanations used in-app by WhatsApp, Viber, and Telegram do not improve user understanding. Other explanations can be better than these current ones.

From an experiment design and metaphor design perspective, we have confidence that our proposed method is better than published alternatives. The design allows for three points of constraining an acceptable metaphor – user descriptions, criteria-based analysis, and survey of user understanding. This style of carefully adding complementary constraints is similar to how scientists arrive at adequate explanations for complex phenomena [40]. Through these different kinds of constraints and measurements, our method provides potential explanations to why metaphors succeed or fail. These potential explanations are in turn candidates for future studies and improvements.

In considering principles for studying security and privacy from a user perspective, Krol et al. recommend – among other things – to "think carefully about how meaning is assigned to the terms threat model, security, privacy, and usability" [41, p. 1]. Our work most clearly focuses on the meaning of security and privacy. As noted in Section IV, studying ways to cue accurate threat models in users is a potential direction for future work. These are the attributes on which our survey tests understanding. Based on the survey, end-users do not appear to have the same use of the concepts as the technology provides. The industry jargon-based descriptions have no impact when mistakes are not especially punished, but do show a significant decrease in the score of participant understanding when such negative changes receive a higher punishment. This result

implies that the metaphors are not improving understanding of risk, but rather just moving it around.

It is unsurprising that if users do not have the same meaning assigned to these terms as the developers do, there will be problems. The user interviews support this claim that users in fact do not use the words the same way. For E2E-encryption generally, (1) participants do not know the meaning of the system, (2) they give incorrect non-technical descriptions of the system, and (3) they give incomplete non-technical descriptions of the system. This state of participant understanding helps explain why the metaphors based on technical jargon and structural mental models do not perform well.

The most natural question is to ask why users do not have the correct understanding. From the data we have available here, there are a number of avenues which can be explored. For one, other work has suggested that security concepts are too complex to explain in a simple metaphor, certainly for the purposes of risk communication [42]. Future work may explore ways to address specific elements of how E2E-encryption works which require functional contribution from users; that is, focus on explaining correct user behaviours to enact, as opposed to requiring users understand the system structure correctly. This can be put into action by developing task-action statements to test metaphors against when applying our rigorous methodology.

Complementary to this, our results strongly suggest that metaphors for security communication must be rigorously tested. We found that existing explanations created by domain experts fall short due to their attempt to engender structural mental models, and this is not doable in the available space and time. In addition, Sasse et al. [43] argue that the goal of effective security can be achieved by considering the primary task, context of use, as well as strengths and weaknesses of users. These factors are still widely ignored even when designers attempt to incorporate usability; they ultimately develop systems based on their perception of what is usable, where it may be more productive to focus on the mental models and capabilities of users [44].

Furthermore, both new and old metaphors cause measurable harm to participant understanding. This harm appears evenly distributed whether participant beliefs start off accurate or not, and appears independent of participant experience with the technology. Related works examining metaphors for E2E-encryption (e.g., [13]) have not included in their methodology design the capacity to check for negative impact upon participants' prior beliefs. This omission creates a clear opportunity for bias, in which researchers selectively measure positive changes without checking what the collateral damage to understanding has been. Testing for harm as well as improvement is a clear contribution of our own approach which we recommend be adopted in future studies.

When asked about WhatsApp's explanation of E2E-encryption, interview participants offered comments such as:

> "Because on WhatsApp...when you start chatting it's just...they just [said] that it's an end to end encryption but it's not really like they give more explanation on how they're doing it basically. Just like talking to people and at the end said that it's like that" (emphasis added).

Such statements corroborate and make concrete the failure to improve user understanding that is evidenced by the survey.

Another point worth noting is that how well a metaphor does depends on how much weight we assign to conceptual baggage. This impact is particularly clear with Statement 4 (If somebody hacks your phone, they will be able to read your messages). The WhatsApp explanation especially makes people falsely believe their messages are secure at rest because they have been encrypted in transit. Advertisers have an incentive to emphasise the positive aspects and downplay short-comings. A balanced understanding of the usefulness of a technology requires just that — balance. However, we hypothesise that one piece of conceptual baggage we identify for the WhatsApp metaphor — the identification of E2E-encryption with the whole of the WhatsApp software package — is a candidate for explaining why this metaphor performed so poorly on Statement 4. This hypothesis is an example candidate for follow up work.

B. General Discussion

Transferring knowledge or understanding to novices is an important aspect of usable security. We can view the task of devising explanatory metaphors for E2E-encryption via work on resituating knowledge more generally [45]. The knowledge is local to experts, and the target audience is novices. The jargon-based explanations essentially try to bridge the knowledge directly to the novices. The more friendly metaphors serve as an intermediate generalisation which the novice can then attempt to localise to their personal situation. Our method of both analytic and survey-based evaluation of metaphors contributes to methods for evaluating when a generalisation is justified. Other methods of justifying a generalisation might use the entities involved and what they do [40] or details of how the knowledge was generated [46]. These modes of scientific explanation can be considered analogous to functional mental models, as opposed to the structural models more popularly associated with science.

The question of finding an adequate metaphor for explaining encryption properties is not merely poetic. Humans' conceptual system is largely based on metaphor, and *"the way we think, what we experience, and what we do every day is very much a matter of metaphor"* [47, p. 3]. Therefore, orienting novices with a helpful metaphor for E2E-encryption should positively impact their use of the technology. An area for future work includes testing this assertion, assuming a metaphor that cues improved understanding can be found. However, our results suggest that – even with improved methodologies for testing metaphors – identifying a *helpful* metaphor may be difficult.

Communicating complex models is a problem among sciences as well as from experts to novices. Therefore, another area for future work would be to test whether explanatory strategies from a science of information security cue user understanding. A pluralist, mechanistic explanation style is better suited to cue user understanding than the laws-based approach often associated with science by security researchers [46]. Our study is a part of a larger discourse on adequate ways to communicate expert knowledge effectively. In this context, the potential uses of our study method both advance beyond usable

security and also, within usable security, go beyond messaging apps that use E2E-encryption.

C. Limitations

Although we have used interviews with end-users to seed our model of participant understanding, we have not directly queried our survey participants about how they explain the topic. Therefore, we do not have direct reports that could inform why the participants answered the survey the way they did. The secondary analysis of the interviews found that interviewees struggled to articulate their understanding of E2E-encryption (which is itself jargon-heavy). One natural avenue for future work would however be to extend the survey, and prompt survey participants to explain their reasoning about the statements in their own words and how this was influenced by the metaphors.

VII. Conclusion

We have contributed a method for evaluating explanatory metaphors in usable security. We have tested this method by evaluating existing descriptions and potential metaphors for E2E-encryption. Our evaluation finds that existing metaphors do not appear to increase user understanding of the functionality of E2E-encryption, though they do appear to shift misunderstanding from one area of functionality to another. We have found we can design new metaphors based on user interviews that have less conceptual baggage than those based on technical jargon. The focus of our new metaphors aims to avoid the dominant approach based on cuing structural mental models.

The results for even our carefully crafted metaphors were negative; they do not improve participant responses. Some of our new metaphors – those which were generated based on participant interviews – perform better, in that they harm participant understanding less than the existing metaphors. This result indicates structural mental models are likely the wrong goal when explaining complex systems to users. Future work would likely have more success if it focuses on cuing functional (i.e., task-action) mental models. Furthermore, when building and testing metaphors, the community should measure the harm to participant understanding as well as any improvements.

Acknowledgements

Demjaha is supported through a Doctoral Studentship (TU/C/000026) granted by the Alan Turing Institute. Spring is supported by University College London's Overseas Research Scholarship and Graduate Research Scholarship, orcid.org/0000-0001-9356-219X (Spring).

Dataset

The data and analysis methods from our survey are available at DOI 10.14324/000.ds.10042529.

References

[1] J. Katz and Y. Lindell, Introduction to modern cryptography. CRC press, 2014.

[2] A. Whitten and J. D. Tygar, "Why Johnny can't encrypt: a usability evaluation of PGP 5.0," in *USENIX Security Symposium*, 1999.

[3] S. Fahl, M. Harbach, T. Muders, M. Smith, and U. Sander, "Helping Johnny 2.0 to encrypt his Facebook conversations," in *SOUPS*, ACM, 2012.

[4] W. Bai, M. Namara, Y. Qian, P. G. Kelley, M. L. Mazurek, and D. Kim, "An inconvenient trust: user attitudes toward security and usability tradeoffs for key-directory encryption systems," in *SOUPS*, ACM, 2016.

[5] A. Herzberg and H. Leibowitz, "Can Johnny finally encrypt? Evaluating E2E encryption in popular im applications," in *ACM Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, 2016.

[6] R. Abu-Salma, K. Krol, S. Parkin, V. Koh, K. Kwan, J. Mahboob, Z. Traboulsi, and M. A. Sasse, "The security blanket of the chat world: an analytic evaluation and a user study of telegram," in *EuroUSEC*, Internet Society, 2017.

[7] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith, "Obstacles to the adoption of secure communication tools," in *IEEE Symposium on Security and Privacy*, IEEE Computer Society, 2017.

[8] J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975.

[9] F. Halasz and T. P. Moran, "Analogy considered harmful," in *CHI '82*, Gaithersburg, Maryland, USA: ACM, 1982, pp. 383–386. DOI: 10.1145/800049.801816.

[10] L. Clark and M. A. Sasse, "Conceptual design reconsidered: the case of the internet session directory tool," in *People and Computers XII: Proceedings of HCI*, vol. 97, 1997, pp. 67–84.

[11] A. diSessa, "Models of computation," *User Centered System Design: New Perspectives on Human Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum*, 1986.

[12] R. M. Young, "Surrogates and mappings: two kinds of conceptual models for interactive devices," *Mental models*, vol. 37, pp. 35–52, 1983.

[13] A. Whitten, "Making security usable," *Unpublished Ph. D. thesis, CS, CMU*, 2004.

[14] J. L. Alty, R. P. Knott, B. Anderson, and M. Smyth, "A framework for engineering metaphor at the user interface," *Interacting with computers*, vol. 13, no. 2, pp. 301–322, 2000.

[15] S. L. Garfinkel and R. C. Miller, "Johnny 2: a user test of key continuity management with S/MIME and Outlook Express," in *SOUPS*, ACM, 2005, pp. 13–24.

[16] S. Sheng, L. Broderick, C. A. Koranda, and J. J. Hyland, "Why Johnny still can't encrypt: evaluating the usability of email encryption software," in *SOUPS*, 2006.

[17] S. Ruoti, N. Kim, B. Burgon, T. Van Der Horst, and K. Seamons, "Confused Johnny: when automatic encryption leads to confusion and mistakes," in *SOUPS*, ACM, 2013, p. 5.

[18] S. Ruoti, J. Andersen, D. Zappala, and K. Seamons, "Why Johnny still, still can't encrypt: evaluating the

usability of a modern PGP client," *arXiv preprint arXiv:1510.08555*, 2015.

[19] V. Roth, T. Straub, and K. Richter, "Security and usability engineering with particular attention to electronic mail," *International Journal of Human-Computer Studies*, vol. 63, no. 1, pp. 51–73, 2005.

[20] W. Tong, S. Gold, S. Gichohi, M. Roman, and J. Frankle, "Why king george iii can encrypt," *Freedom to Tinker*, 2014.

[21] J. Lausch, O. Wiese, and V. Roth, "What is a secure email?" In *Proocedings of EuroUSEC'17*, Internet Society, vol. 2, 2017.

[22] P. N. Johnson-Laird, Mental models: Towards a cognitive science of language, inference, and consciousness, 6. Harvard University Press, 1983.

[23] M. A. Sasse, "Eliciting and describing users' models of computer systems," PhD thesis, University of Birmingham, 1997.

[24] K. Renaud, M. Volkamer, and A. Renkema-Padmos, "Why doesn't jane protect her privacy?" In *International Symposium on Privacy Enhancing Technologies Symposium*, Springer, 2014, pp. 244–262.

[25] R. Wash, "Folk models of home computer security," in *SOUPS*, ACM, 2010, p. 11.

[26] K. Romer, Episode 773: slot flaw scofflaws: planet money: npr, http://www.npr.org/sections/money/2017/05/24/529865107/episode-773-slot-flaw-scofflaws, (Accessed on 09/04/2017), 2017.

[27] R. P. Articles, The ceremony - radiolab, http://www.radiolab.org/story/ceremony/, (Accessed on 09/04/2017), 2017.

[28] Analogies archive - the analogies project, https://theanalogiesproject.org/the-analogies/, (Accessed on 09/04/2017).

[29] O. L. Dictionaries, Analogy - definition of analogy in English — Oxford Dictionaries, https://en.oxforddictionaries.com/definition/analogy, (Accessed on 09/02/2017).

[30] B. Anderson, M. Smyth, R. P. Knott, M. Bergan, J. Bergan, and J. L. Alty, "Minimising conceptual baggage: making choices about metaphor," in *BCS HCI*, 1994, pp. 179–194.

[31] S. Brostoff, M. A. Sasse, D. Chadwick, J. Cunningham, U. Mbanaso, and S. Otenko, "'r-what?' development of a role-based access control policy-writing tool for e-scientists," *Software: Practice and Experience*, vol. 35, no. 9, pp. 835–856, 2005.

[32] R. W. Shirey, "Internet security glossary, version 2," 2007.

[33] E. Barker, "Recommendation for key management–part 1: general (rev 4)," U.S. National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. SP 800-57, 2016.

[34] R. Ross, M. McEvilley, and J. C. Oren, "Systems security engineering: considerations for a multidisciplinary approach in the engineering of trustworthy secure systems," U.S. National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. SP 800-160, 2016.

[35] S. Parkin, A. Boscor, K. Y. Cheng, K. Karunananthan, and G. J. Tang, "Talking your way out of trouble: designing for trust in WhatsApp encryption features," in *under review*, 2019.

[36] K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.

[37] B. Anderson and J. L. Alty, "Everyday theories, cognitive anthropology and user-centred system design," in *BCS HCI*, 1995, pp. 121–135.

[38] I. Becker, S. Parkin, and M. A. Sasse, "Measuring the Success of Context-Aware Security Behaviour Surveys," in *Learning from Authoritative Security Experiment Results (LASER)*, Arlington, VA: USENIX Association, 2017.

[39] N. Thayer-Hart, J. Dykema, K. Elver, N. C. Schaeffer, and J. Stevenson, Survey fundamentals: A guide to designing and implementing surveys. University of Wisconsin, 2010.

[40] J. M. Spring and P. Illari, "Mechanisms and generality in information security," *Submitted Manuscript*, 2017.

[41] K. Krol, J. M. Spring, S. Parkin, and M. A. Sasse, "Towards robust experimental design for user studies in security and privacy," in *Learning from Authoritative Security Experiment Results (LASER)*, IEEE, San Jose, CA, 2016, pp. 21–31.

[42] L. J. Camp, "Mental models of privacy and security," *IEEE Technology and society magazine*, vol. 28, no. 3, 2009.

[43] M. A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'weakest link'–a human/computer interaction approach to usable and effective security," *BT technology journal*, vol. 19, no. 3, pp. 122–131, 2001.

[44] L. F. Cranor and S. Garfinkel, Security and usability: designing secure systems that people can use. O'Reilly Media, Inc., 2005.

[45] M. S. Morgan, "Resituating knowledge: generic strategies and case studies," *Philosophy of Science*, vol. 81, no. 5, pp. 1012–1024, 2014.

[46] J. M. Spring, T. Moore, and D. Pym, "Practicing a science of security: a philosophy of science perspective," in *New Security Paradigms Workshop*, Santa Cruz, CA, USA, 2017.

[47] G. Lakoff and M. Johnson, Metaphors we live by. Chicago, IL: University of Chicago Press, 1980.