"If you were attacked, you'd be sorry": Counterfactuals as security arguments

Cormac Herley*
Microsoft Research
Redmond, WA, USA
cormac@microsoft.com

Wolter Pieters
University of Twente & TU Delft
Enschede / Delft, Netherlands
w.pieters@tudelft.nl

ABSTRACT

Counterfactuals (or what-if scenarios) are often employed as security arguments, but the dos and don'ts of their use are poorly understood. They are useful to discuss vulnerability of systems under threats that haven't yet materialized, but they can also be used to justify investment in obscure controls. In this paper, we shed light on the role of counterfactuals in security, and present conditions under which counterfactuals are legitimate arguments, linked to the exclusion or inclusion of the threat environment in security metrics. We provide a new paradigm for security reasoning by deriving essential questions to ask in order to decide on the acceptability of specific counterfactuals as security arguments, which can serve as a basis for further study in this field. We conclude that counterfactuals are a necessary evil in security, which should be carefully controlled.

CCS Concepts

•**Security and privacy** → **Social aspects of security and privacy;** *Logic and verification; Economics of security and privacy;*

Keywords

adversarial risk, control strength, counterfactuals, security arguments, security metrics, threat environment

1. INTRODUCTION

The observation that absence of attacks doesn't imply security is commonplace. We can't use the fact that a server that stores passwords in plaintext goes years without incident to argue that there is no risk. If an electronic voting system isn't attacked (or if there is no evidence of attacks), this doesn't mean that the system is secure.

Clearly there is a need when designing systems that will be used in adversarial environments to consider what might happen, not just what has been observed to happen. Therefore, there is an obvious need to engage in counterfactuals, or what-if scenarios, when discussing security. Such counterfactuals would then state that if, contrary to reality, attacks *would* take place, the systems under consideration would or would not be secure enough. However, such arguments can also be employed to emphasize the need for obscure controls, based on possible threats that may not be very likely. This illustrates that it is largely unclear under which conditions counterfactuals justify the consumption of defensive effort.

To be a bit more precise, counterfactuals are used to speak about events in conditions that are not real. This may be either because we consider something that is contrary to the actual course of events in the past ("if I had attacked you"), or because we consider something that might be the case in the future ("if I would attack you"). In both cases, the occurrence of the threat is what is assumed contrary to reality, or as a possible reality. We will refer to both of these instances as counterfactuals, hypotheticals, or what-if scenarios, as the time difference is not our primary concern here. We will make more precise distinctions, for the security domain only, later in this paper.

The use of counterfactuals is connected to the adversarial threats considered in security research. The ability to "think like an attacker" is prized among security researchers. Spotting a new vulnerability or problem often guarantees a publication at a prestigious academic security conference. Presentations at non-academic conferences, such as Blackhat and Defcon, are dominated by demonstrations of previously unknown exploits. Debates on national security and terrorism issues often justify measures, based not on what has been observed, but on speculation about what might happen. Schneier describes [1, (March 25, 2008)] "the security mindset involves thinking about how things can be made to fail. It involves thinking like an attacker, an adversary or a criminal." A common line of reasoning says that a *sufficiently motivated attacker* will always find a way in. A popular textbook [26] effectively argues that all possibilities must be considered:

> *Principle of easiest penetration: an intruder must be expected to use any available means of penetration.*

This paradigm of considering all possibilities leads to many threat scenarios which might result in harm that we currently don't see. For example, Aviv et al. [3] report that the smudge marks left when a user unlocks a touchscreen device *can* be used to assist an attacker guess the PIN. Backes et

*Authors listed in alphabetical order of last names.

NSPW '15, September 08 - 11, 2015, Twente, Netherlands

ISBN 978-1-4503-3754-0/15/09…$15.00

DOI: http://dx.doi.org/10.1145/2841113.2841122

al. [22] describe how the image of an LCD screen, reflected on a shiny surface *can* be read from a distance of 20m using a telescope. Koscher et al. [19] describe how an adversary who has access to the debug port of a modern car *can* alter it's behavior in unsafe ways (e.g. by interfering with the brakes). But of course not everything that can happen does happen. Are these realistic threats or not? Do people suffer harm because of these threats and is spending resources to protect against them advisable, or are they scenarios that are possible in theory but represent little actual threat?

It seems clear that we can think of an unlimited number of scenarios, ranging from those that argue for steps as simple as password masking, to ticking time bomb scenarios that attempt to justify torture. These what-if scenarios make claims on what could happen, but they also make claims about what needs to be done, which depends heavily on the context. It has been shown that mobile banking malware is possible [9], but still the banks claim the app is the preferred channel (and has much fewer incidents than the website). Clearly some counterfactuals are more realistic than others, but how can we decide?

If resources were unlimited, or countermeasures were costless, then of course we could take action against every possible scenario we could think of (and could distinguish from legitimate behaviour). However, when resources are finite tradeoffs must be made. If we cannot defend against every possible threat scenario then decisions must be made. The resources asked by some what-if scenarios must be denied so that others can be granted. It is simply not possible to invest defensive effort in response to all of them. Lampson summarizes the situation succinctly [20]:

> *Security experts always have a plausible scenario that demands a new option, and a plausible threat that demands a new defense.*

Thus the question of whether to invest or not is unavoidable. On what basis do we decide that some scenarios are plausible and some others ridiculous? Using a soft criterion like "ridiculous" seems to ensure that resource allocation decision will be made in an unsystematic way. Few argue that perfect security is attainable. Even those who caution that all eventualities must be considered will probably acknowledge that some scenarios are too far-fetched to merit countermeasures and there must be a limit to spending. However, while everyone acknowledges that there must be a limit to the scenarios we consider nowhere are there explicit guidelines about what can be neglected.

This paper aims at making the role of counterfactuals as security arguments explicit. In Section 2, we provide examples of the use of counterfactuals in security reasoning. In Section 3, we discuss how counterfactuals are used when measuring security. In Section 4, we discuss the role of counterfactuals in security risk management, notably in terms of the differences between (i) likelihood of occurrence and likelihood of success, and (ii) fault trees and attack trees. In Section 5, we provide conditions under which counterfactual security arguments are acceptable, and we conclude in Section 6.

2. COUNTERFACTUALS IN CASES

We begin with a few more examples of counterfactuals, illustrating different aspects of their necessity as well as their problems. We seek to show the magnitude and importance of the open problem caused by the lack of tools to reason about hypothetical threat scenarios.

2.1 Password masking

Most programs and web services mask password characters as they are entered by the user, presumably to guard against visual collection. Of course, most of the time that passwords are entered there is no threat, and humans are good are noticing other humans in their proximity.

The argument for password masking is that *if* someone attempts shoulder surfing, masking the characters complicates the task. An argument against is that we have no idea how frequently shoulder surfing happens (if at all) so it is hard to argue that the benefits outweigh the costs: the attack is non-scalable [14] and a determined attacker can acquire the password at lower cost.

Significant difference of opinion emerged when security expert Schneier declared that he thought the practice unnecessary only to reverse himself days later (after considerable argument in the blogosphere) [1, (July 3, 2009)]: "So was I wrong? Maybe. OK, probably."

How might we decide the question as to whether password masking makes a difference? One approach might be to do a randomized trial: switch off masking for a random subset of users at a large web-site for a period and observe whether hijacking rates differ between the treatment and no-treatment cases. If we observe a statistically significant difference then masking clearly has an effect (and we might then discuss whether the effect justifies the cost). However, in the absence of a significant effect can we conclude that the measure is worthless? It might be that the sample studied does not have the statistical power to reveal the difference. The fact that the absence of improved outcomes does not indicate that the scenario can be ignored is shown in our next example.

This case illustrates that the difficulty of testing the efficacy of security measures. Without relying on counterfactuals, tests of effectiveness of security controls would always have to wait until a threat materializes in the real world. If one does not want to wait, one tests instead how the control behaves under the threat against which it is meant to protect. We will come back to what the outcomes mean in the next section.

In principle this case should be simple. Whether masking is worthwhile or not is a more-or-less binary decision. The fact that there is no agreement on the question appears in part due to the absence of data documenting the frequency of the threat, which may be either due to the threat not occurring (which can change), or lack of measurements of a threat that does occur.

2.2 Lifeboats

Still, frequencies may not always be the basis for a decision. If we did a randomized trial of passenger ferries where half had lifeboats and half did not the expected result is no observable difference in outcomes for the passengers involved. Passenger ferries sink very rarely. For example, the Washington State ferry system in the US carries about 12 million passengers per year and has not had a sinking since its foundation in 1951. A half century-long randomized trial would conclude that lifeboats have no observable effect.

Yet, clearly we are reluctant to draw the conclusion that lifeboats are unnecessary. *If* a ship sinks, lifeboats prevent

loss of life. In cases such as that of the Titanic, making it onto a lifeboat or not made the difference between life and death. That offers one clear point of contrast with the password masking question: while sinkings might be rare the consequences can be catastrophic, so even if we consider the counterfactual scenario very unlikely, the cost of the measure to address it does not seem worth saving.

This case illustrates that counterfactuals have an important role in safety and security engineering. It shows that if a particular threat materializes, harm can be mitigated with suitable controls. Simply arguing that the threat is not likely can be dangerous. On the other hand, applying all kinds of controls simply because a what-if argument exists is also inefficient.

2.3 e-voting

In the Dutch e-voting controversy [17], there was discussion on whether the machines used until 2007 were secure enough. The ministry claimed the machines were secure enough for the Dutch context; a pressure group asked whether the machines would also be secure enough if they were deployed in <country-at-war>. The machines were never deployed in <country-at-war>, so there is a legitimate question here on the relevance of this argument. Similarly, one can question the ministry's assessment, as the Dutch context could also be subject to change, and "past performance is no guarantee of future results". The fact that no incidents or losses occurred can be a result of good security, but also of lack of attacks (or even of well-hidden attacks).

This case illustrates the role that (changing) threat environments play in the discussion on counterfactuals. If we conclude that the e-voting machines aren't secure enough for <country-at-war>, probably because of a more severe threat environment, then what does that mean for their suitability for the Dutch context? Another illustration of this point might be to consider if a Edward Snowden and an ordinary member of the public had the same password. What might be perfectly adequate to protect someone who is not the subject of targeted attack would be completely ineffective against a well-resourced and determined adversary. The context clearly makes a large difference to the likelihood of attack. Even the context can change unpredictably and without warning: the email account of Tryvon Martin, the Florida teen killed in controversial circumstances in 2012, was hacked. An email account of a teenager, which would ordinarily be a low-value target, suddenly became a high-value on account of the circumstances of the death of the owner.

2.4 Cockpit security

A recent case where security controls played a controversial role is the Germanwings crash. Without claiming a definitive cause ahead of official investigations, let's assume that the crash was indeed caused deliberately by the co-pilot after the pilot left the cockpit, that he denied access to the pilot, and that the pilot was not able to force his way in.

Let's also assume that certain controls were implemented based on certain security arguments (without claiming any resemblance to the real course of events). The arguments could look like this: "If a terrorist manages to enter the cockpit, we have a problem." And: "If a terrorist tries to enter the cockpit, a strengthened door and the ability to deny access completely will be effective."

Again, this case illustrates the inevitability as well as the danger of engaging in counterfactuals. In particular, this case shows that the acceptability of a counterfactual argument cannot be determined from a single threat scenario or a single control only. The control may have adverse effects in case of different threat scenarios, or may only work (or not work) in combination with other controls. In particular, different threat scenarios may involve different types of attackers, such as outsiders versus insiders.

This example is interesting in that it illuminates a two-sided nature to the problem. For many scenarios the suggested countermeasure reduces risk, but there is a question as to whether the risk is real. Here, the countermeasure reduces one risk while increasing another. That is, a reinforced cockpit door reduces risk if the danger is an attacker on the cabin side of the door; however, it increases the risk if an attacker is already in the cockpit, or everyone in the cockpit is disabled. Thus, the question of under what circumstances a cockpit door can be opened from the cabin side is not simply an evaluation of the likelihood of a particular scenario, but a judgement of the relative likelihoods of several.

2.5 Analysis

The counterfactuals we have presented illustrate certain of the difficulties of reasoning in this space. A reasonable starting point would be to assume that a rational defender deploys a countermeasure if the cost of a countermeasure is less than the expected loss that it prevents. For simple predictable phenomena (such as shoplifting at a major store chain) we might be able to write this as

$$p \cdot L > C, \tag{1}$$

where p is the probability of an attack succeeding, L the loss incurred and C the cost of the countermeasure. (These may not be objective, and different people may have different estimates.)

The what-if scenarios we have examined illustrate several complications that rule out applying such a simplistic analysis.

- Rare or unlikely events with catastrophic consequences (lifeboats, cockpit door)
- Active adversary (e-voting, cockpit door)
- Countermeasures do not align with our values (torture or abrogation of voting rights)
- Is a countermeasure always appropriate, or only under certain circumstances? (e-voting)
- Countermeasures reduce one, but increase other risks (cockpit door).

These factors increase the difficulty of examining countermeasures from the viewpoint of costs and benefits. Rare events with catastrophic outcomes represent a challenge; as $p \to 0$ and $L \to \infty$ then $p \cdot L$ is extremely hard to determine accurately. The counterfactual makes a statement about the future, but does not commit to any time interval. The statement that something bad will happen within a year is testable and concrete while the claim that something bad will happen is not [15]. Active adversaries represent a challenge, since usually we estimate likelihoods based on past

events; an active adversary can present us with an attack that has never occurred before. This presents difficulty for the approach of representing attacks as probabilities. Countermeasures that do not align with our values have costs that are very hard to evaluate; it is exceedingly difficult to use cost-benefit analysis when discussing rights that we consider inalienable (e.g. voting, freedom, privacy). Countermeasures that apply some of the time and not others suggest that the $p \cdot L > C$ decision must be done separately in each of the identifiable sub-problems. Finally, countermeasures that reduce one but increase another risk, suggest that $p \cdot L$ must be split apart into several components.

Given these arguments, it seems clear that a general approach to determine whether a what-if scenario justifies a proposed countermeasure will be very hard to get. Uncertainty about any of the factors in the cost equation makes this task difficult, and yet in many situations we have uncertainty about all of the factors, and little hope that better information will become available. That is, decisions must be made under uncertainty. This makes it look like the central question is where the burden of proof should lie. Do we spend by default unless a scenario can be shown not to be a threat, or do we withhold by default unless the scenario has demonstrated realistic ability to harm?

We will investigate these themes in the following subsections. We start with the question of testing / metrics. Then we'll investigate counterfactuals in risk management, which is related to the topic of threat environments. We will come back to the issue of burden of proof in the discussion at the end of this paper.

3. COUNTERFACTUALS IN METRICS

Counterfactuals are not only apparent from concrete examples in security. They are also deeply embedded in the methodologies of security research. We will show this from the perspective of security metrics (this section) and security risk management (next section).

Essentially, security metrics aim at providing quantitative statements on how secure something is. There are different ways to do this. A secure neighbourhood might be defined as a neighbourhood with a low incidence of crimes. A secure house might be defined as a house that is difficult to get access to (without a key). We have observed in the Dagstuhl seminar on socio-technical security metrics [13] that many core disagreements and misunderstandings regarding security metrics can be traced back to a focus on counterfactual metrics as compared to non-counterfactual ones. We will explain the embedding of counterfactuals in security metrics next, by illustrating different types of security metrics and explaining their relation to counterfactuals.

3.1 Incident counts

As we discussed, a seemingly easy way to avoid counterfactuals altogether is to look at frequencies of actual threats. Many security metrics are therefore based on counting incidents, for example in terms of infected machines of Internet service providers [36]. If a provider has fewer infected machines (relative to its size), then it would be more secure. When also the impact of the incidents is included, one can sum up the impact of all the incidents to obtain the overall loss (assuming the loss is expressed in terms of money). A system A would then be more secure than system B if the average loss (per unit of time) is lower. When extrapolating to the future, this metric is called annual loss expectancy. These metrics do not use counterfactuals, but are based on observed incidents (facts). This also means that the metrics include the actual behaviour of the threat agents (attackers): if the attackers would not be active, there would be no incidents and no loss.

3.2 Penetration testing

Metrics used in penetration testing are of an entirely different category. In penetration testing, professional testers aim at gaining access to an organisation's assets. This may include digital hacking, but also physical trespassing and social engineering. One can then measure for example success, but also the time taken by the attackers. As these are not real attackers, they basically impersonate the counterfactual argument: if I would attack you, how far would I get (and how quickly [2])? One cannot count these as real incidents, as the attacks were planned by those who were trying to measure security.

This is also related to the notion of control: if one can control the threat environment, the results that one obtains become *independent* of the threat environment. A small analogy: the truth of the statement "ice melts" is dependent on the environment, whereas "ice melts at 293 K" is independent of the environment, because it makes the environment explicit. The latter can also be formulated as "if I apply a temperature of 293 K, ice will melt". Thus, paradoxically, by making the threat environment *explicit* in the metrics, the metrics become *independent* of the *actual* threat environment (but at the same time counterfactual).

An excellent example of making the threat explicit is found in physical security metrics. *Burglar resistance* indicates how difficult it is for a burglar to open a door or window by force (European standards EN 1627 – 1630). Different levels of burglar resistance are defined in terms of the time required to enter for a *specific type of adversary with specified tools*. For example, burglar resistance class 2 means at least 3 minutes delay for a burglar with screwdrivers, a hammer, etc. These times are tested in a laboratory by professional testers with the specified equipment. Note that these metrics do not claim anything about the modus operandi of real burglars, or of the value of the assets protected by the resistance classes. Therefore, they are what-if metrics: under the specified threat conditions, we give these guarantees.

For burglar resistance, there are (implicit) assumptions on interpolation and extrapolation. If we give the burglar slightly better tools, he will probably be slightly faster (and not slower; the relation is monotonic). This points to another important lesson: what-if scenarios, especially in the context of metrics, require (explicit or implicit) assumptions on the result if the scenario would be (slightly) different.

3.3 Correlation versus causation

We have mostly been talking about counterfactuals in terms of *threat* up to now. This is the title of the paper: "If you were attacked, ...", or "If I had attacked you, ...". However, there appears to be a second type of counterfactual, which is related to controls rather than threats. It goes like this "If you had implemented this control, ...". Of course, both types can be combined: "If I had attacked you, and you had implemented this control, ..."

In order to understand the control-type counterfactual better, we have to dig into the discussion on effectiveness

of controls. What does it mean if we observe that systems that have a control are more secure than the ones that don't? For example, computers with antivirus installed are infected less often than ones without? This doesn't necessarily mean that the improvement can be ascribed to / is caused by the control. If computers that have antivirus are infected less often, this may be because users who install antivirus are more security-aware, and therefore less likely to get infected, even if the antivirus is useless. So, in real-world circumstances, without counterfactuals, it is hard to say something about the effect of controls.

This is linked to the need for controlled experimental approaches instead of observations. If, in an experimental setting, we randomly assign the experimental and the control condition (with and without antivirus), we avoid self-selection bias. The counterfactual here lies in the fact that rather than using real-world selection of antivirus by users, we ask how likely computers are to get infected *if they would have antivirus* (or not). The counterfactual lies in the random assignment (we interfere in the world, and without the experiment the actual (factual) situation would have been different).

In the medical domain, this would be related to randomly applying vaccination and control conditions, and then measuring infection rates in the real world. One can then make claims like "if you would apply this treatment" rather than "patients who received this treatment". Again, in the latter case, the effect may be caused by something else than the treatment, such as placebo effects.

3.4 Controlling for the threat environment / attacker behaviour

Controlling the condition may not be enough though. If we want to check the effect of a control, and the threat never materialises during the experiment, we learn very little. In the medical domain, one limits the subjects to patients only, but in the cyber domain, with *preventive controls*, this does not make much sense. One does not know who will be affected by the threat and who won't. Therefore, it may be necessary to control the threat as well. This is for example the case in phishing [10] and social engineering experiments [6], where scripted attacks are executed after the participants have been subjected to the experimental or the control condition.

In the medical domain, the experimental approach would be comparable to randomly applying vaccination and control conditions, and then measuring infection rates *after injecting the pathogens* that the vaccination is supposed to protect against. (This is all purely hypothetical.) This would provide statements including both threat and control: "If one would receive vaccination, and if one would then come in contact with the pathogen, ..." Analogously, in security, one can test the effect of controls only if the threat occurs, which provides a reason for administering the threat artificially.

Although controlling the threat environment in an experimental setting makes it more likely that security differences show effects, it leads to another problem. For how do we know that the threat environment created in the experiment is similar to the threat environment in the real world? In the study, this only happens in controlled settings, so one cannot make claims that relate to reality. If the pathogen does not occur outside the lab, the whole exercise is pointless. (Unless one takes into account the possibility of accidental release, which would correspond to attacks being used in the real world that were devised in security experiments.) And, once we know how a system responds to controlled threats, how can we extrapolate that knowledge to other threats? In the medical setting we probably know that the pathogen is out there when developing vaccination, but threats may be more dynamic in security (although pathogens may also mutate).

3.5 Analysis

Security metrics can be based on resistance or on incidents. In the first case, the threat environment is excluded from the metrics, by means of providing an *artificial* threat against which the resistance can be measured (e.g. in terms of time needed to get in). In the second case, the incidents are determined by both the resistance (also called security level [5]) and the threat environment. We have termed these type 1 and type 2 security metrics respectively [13]. Type 1 metrics are necessarily counterfactuals, as they do not rely on the threat environment in the real world. Therefore, with a type 1 metric, I can say that it is very easy to eavesdrop on passwords if they aren't masked. With a type 2 metric, I can say that even so, passwords are very rarely stolen this way in the real world.

Type 1 metrics are useful for comparing system A against system B under the same threat environment, while their threat environments are different in reality. For example if one wants to know whether the e-voting system in use in <country-at-peace> is more or less secure than the system in use in <country-at-war>. Results of penetration tests, controlling the threat environment, may provide the necessary counterfactual arguments here. If one would compare the systems by counting incidents, this would obviously be unfair.

If one is interested in questions of, for example, return on security investment [7], then type 1 metrics (counterfactuals) are not so useful. If a system is not so secure, but there is no incentive for attackers to attack it, then it does not make much sense to invest in security. This is similar in structure to the Titanic argument: there is no threat, so we don't need controls.

Type 1 metrics are also useful for low probability / high impact events (black swans). There may not be enough data on sunken Titanic-style ships, but we may still want to make sure that people can survive if something of the sort happens.

Similar arguments can be made for control-type counterfactuals. One can measure the effectiveness of a control in terms of reduction of actual incidents, or by the reduction of success of experimentally administered threats.

It is not obvious which type of metric is more valuable. If somebody claims that upgrading my house from class 2 to class 3 burglar resistance increases the required time for a burglar from 3 to 5 minutes, this may not tell me much about how much less likely I will be to get a burglary if I do. But if one tries to find such statistics, it might as well turn out that class 3 houses have the same chance of burglary, because they are typically the larger houses with higher expected gain for burglars.

4. COUNTERFACTUALS IN RISK

Security risk management comes from a different tradition than security metrics. In particular, risk analysis and risk

management are inescapably bound to notions of likelihood or probability. Rather than defining "success" or "occurrence" as an empirical variable, risk managers talk about things in terms of likelihood of success or likelihood of occurrence, often based on expert judgement.

In addition, because of the uncertainty of future events, risk managers naturally think in terms of what-if scenarios. This means that they are familiar with the use of counterfactuals. Still, in practice, many assumptions on the use of counterfactuals remain implicit, creating a source of misunderstandings.

4.1 Likelihood of occurrence vs. likelihood of success

In security risk management, the confusion around counterfactuals therefore takes a different shape. In particular, the often inconsistent use of the terms "probability" or "likelihood" exemplifies the issue. These can refer to (1) probability or likelihood of occurrence of an event, or (2) probability or likelihood of success of an action or attack step. For example, one can say that a particular area is expected to flood once every 1000 years, or one can say that a phishing attempt has a 20% chance of success. In the former case, they say something about how often an event is expected to occur, and in the second case, they say something about how likely an action is to succeed, *if executed*.

An additional complication is the confusion between probabilities and frequencies in such likelihood statements. If asked whether "once every 1000 years" is a probability, most people will say yes.[1] Only after explaining that one can also say twice per year, then the message comes across: these are not probabilities, but frequencies. They do not have 1 as an upper bound, and their unit is y^{-1}, not 1. Frequencies do have associated probabilities, namely probability distributions that represent the probability of occurrence of an event before time t.

The trick here is that the probabilities (of success), linked to type 1 metrics, represent the counterfactuals. They are probabilities (not bound to time), because they *assume* the occurrence of an event, and try to say something about connected events (linking occurrence to success). On the other hand, the non-counterfactual frequencies (of occurrence), linked to type 2 metrics, represent real-world events. They are frequencies (bound to time), because they say something about how often events occur.

Thus, one materialization of the counterfactuals issue in risk management is the confusion between likelihood of occurrence and likelihood of success. The non-counterfactual approaches focus on likelihood of occurrence (frequency), and the counterfactual approaches on likelihood of success (probability).

4.2 Fault trees vs. attack trees

This is also the fundamental difference between fault trees, used in safety analysis [24], and attack trees, used in security analysis [21, 32]. Both represent connections between events, and associated metrics such as likelihoods. However, fault trees have an undesirable event as root (typically system failure), whereas attack trees have an attacker goal as root. Both roots can be refined using AND- and OR-nodes, representing connections with underlying events / attack subgoals. The key observation from the focus of this paper is that fault trees use probability of *occurrence* of basic events as a standard annotation (typically represented as a cumulative failure probability over time) the latter use probability of *success* of basic attack steps (not bound to time). Therefore, attack trees – in contrast to fault trees – are inherently counterfactual: *if* an attacker would try to reach the goal, with which probability (or time, cost, etc.) would he succeed?

This also means that attack trees and fault trees are incommensurable: one cannot just plug an attack tree into a fault tree as a subtree, or vice versa, an issue that has remained fairly implicit so far. This increases the confusion of likelihood of occurrence and likelihood of success.

4.3 Factor Analysis of Information Risk

In the Factor Analysis of Information Risk (FAIR) taxonomy [35], this has been resolved by calling likelihood of occurrence "Threat Event Frequency" and likelihood of success "Vulnerability". Vulnerability is in turn dependent on properties of the attacker ("Threat Capability") and properties of the defense ("Control Strength").

Together, Threat Event Frequency and Vulnerability determine Loss Event Frequency. In other words, frequency of occurrence and likelihood of success determine frequency of success. The distinction between expected frequencies (occurrence) and probabilities (success given occurrence) helps in resolving the confusion around likelihood. Note that Vulnerability is a counterfactual here: if the threat event would occur, how likely would the threat event be to cause a loss event? Combined with the likelihood/frequency of occurrence of the threat events, this provides a measure of Loss Event Frequency.

In terms of counterfactuals, Vulnerability represents a counterfactual argument or metric: if the threat occurs, how likely is it to succeed / cause damage? By contrast, the Threat Event Frequency represents the threat environment: how often is the threat expected to occur? In adversarial contexts (malicious threats), the Threat Event Frequency may depend on Vulnerability (and expected gain) as perceived by the attacker.

Although FAIR illustrates how one can separate system properties from the threat environment in risk management, it does not provide guidance on when counterfactual arguments are acceptable. Although Vulnerability is inherently counterfactual, and one can compare systems in terms of Vulnerability, this doesn't answer the question whether the e-voting machines (with a certain Vulnerability) are secure enough for the Dutch context. Even if we can show that the Vulnerability is below a certain level, what does that say about acceptability in a particular threat environment?

4.4 Argumentation-based risk analysis

Next to the use of counterfactuals as arguments in general risk assessment, some approaches focus specifically on arguments as the core of risk identification and assessment, such as [16]. In this method, attacker and defender roles are assigned to players, who then try to present and counter arguments for and against security. An argument is valid ("IN") if and only if all its counterarguments have been defeated ("OUT"). Here, the counterfactuals are presented by the attackers: *if* I would do this and this, you'd have a problem. The defenders may provide several counterarguments:

[1] Not a scientific result, speculation from experience only.

you'd never be able to do this, it wouldn't cause any problems if you would, or you'd never want to do this.

This approach illustrates how counterfactuals, rather than being a problem external to risk analysis, can be internalized in the risk management process. This also provides information on how counterfactual security arguments could be refuted. However, also these approaches do not say anything about the acceptability of counterfactual claims, but rather leave this judgement to the players in the argumentation game.

4.5 Analysis

In the domain of risk assessment, counterfactuals appear in notions such as probability/likelihood of success, vulnerability, and resistance. Such concepts say something about the target of analysis under the assumption that certain threats will occur. "If this threat would occur, then ..." The separation between occurrence and success also provides possibilities for disentangling counterfactuals: they may say something about what would happen under certain threat environments, but if such threat environments are unrealistic, the counterfactuals do not make much sense. If the Threat Event Frequency is zero, the Loss Event Frequency will also be zero, no matter how high the Vulnerability.

Whereas risk management is inherently connected to what-ifs, the intricacies of actual or predicted threat environments and counterfactual arguments have only been discussed implicitly in the different frameworks. The problems related to counterfactuals in risk assessment, and the (partial) solution from the FAIR framework, show that likelihood of occurrence and likelihood of success can be separated, where likelihood of success can be tested under artificial conditions (counterfactual), and likelihood of occurrence has to be derived from real-world threat environments. We haven't learnt much about the acceptability criteria for counterfactuals though.

5. CONDITIONS FOR ACCEPTABILITY

There are many reasons for engaging in counterfactual arguments in the security domain, as we have seen in the previous sections, based on cases, metrics, and risk assessment approaches. In this section, we try to identify the most important questions to ask in order to determine whether a specific counterfactual justifies defensive effort.

5.1 When do counterfactuals justify the consumption of resources?

Why is the question whether counterfactuals justify investments so hard? In principle, resources should be spent when the benefit is greater than the cost (e.g. (1) is satisfied). For many of the reasons outlined in Section 2.5 this isn't always possible. In fact, we now concern ourselves only with the case where we cannot determine whether the cost is greater or less than the expected loss. For this to be the case there has to be uncertainty about at least one of the parameters p (probability of success), L (loss incurred) or C (cost of countermeasure). In a majority of the examples studied the greatest uncertainty surrounds p.

One way to argue in favour of a counterfactual claim (an attack scenario that hasn't happened in reality yet) is demonstrating that there is a positive business case for at least one type of adversary; i.e., there are people who would have positive expected utility upon executing the attack scenario, and this utility is higher than (or at least comparable to) that of alternative scenarios. Recall that the cost of attacking must capture all costs, including opportunity costs and those associated with finding profitable targets [11]. If such a positive utility attack case exists then clearly $p > 0$. We propose that attacks which clearly have positive utility to an attacker should be prioritized over those which do not, or for which it is unclear. This approach is very limited if we interpret cost and benefit in a purely monetary sense. For example, Denial of Service or vandalism attacks likely yield little monetary value to the attacker. Thus, attacker utility must be interpreted using knowledge of their motivations. Defacing the website of `whitehouse.gov` has clearly higher utility to many attackers than doing the same to an arbitrary website, even though both acts very likely have low, or zero, monetary value.

A major reason for inability to decide if (1) holds is that we can't estimate the parameters, or the cost of doing so is prohibitively high. For example, for threat events with low probability, population sizes for studies into the effects of controls may become very large. If only one out of ten thousand people is expected to become exposed to a threat event every year, then a sample size in the millions is required to produce significant results. If we consider that the cost of a defensive measure must include the cost of determining whether the measure is worthwhile then we have a way forward. All other things being equal, what-if scenarios that are expensive (or even impossible) to evaluate should have lower priority than ones that are cheap. This is true even if the evaluation has not been performed or planned. Our reasoning is that when the evaluation is expensive it is less likely to be performed; thus, if an ineffective countermeasure is deployed it is less likely to be eventually discovered, and hence has higher expected overall lifetime cost. In this respect we follow Naor [23] who suggests that among competing cryptographic schemes those whose assumptions are most easily evaluated should be preferred. In other words, (1) can serve as a guide even in counterfactual scenarios if we insist that the right-hand side capture all costs including those associated with evaluating the return on investment. Testability is thus a valuable feature in discriminating between various counterfactuals.

5.2 Questions on the acceptability of counterfactuals

Can we leverage the above to make decisions on whether to accept a certain counterfactual security argument? We do this in a two-step approach: first we gather the relevant questions to ask, and then we structure those questions in terms of acceptability criteria.

It is important to distinguish the population we are trying to answer the question for. Is this just a single organisation, or are we addressing the whole population at once? For threat-type counterfactuals, this means if we judge whether the counterfactual threat is relevant for that particular organisation, or for the community as a whole. For the control-type counterfactual, this means if we judge whether the control is effective for that particular organisation, or for the community as a whole. Effectiveness of a control may also depend on the context, for example whether certain other controls are in place. Requiring highly complex passwords is more effective as a control if those passwords are also sent over a secure connection rather than in the clear. This

is because eavesdropping on complex passwords is as easy as eavesdropping on simple ones, so complex passwords are completely ineffective against this threat on insecure connections.

In the questions below, we distinguish between the following items:

- a (counterfactual) threat T
- a counter-measure X that addresses T
- a (defensive) context C

5.2.1 Threat-type counterfactuals

Threat-type counterfactuals are of the form "If I had attacked you (but I didn't), you'd have had a problem (because you don't have the right controls)." Whether we spend based on such counterfactuals depends on the question whether the threat can be determined to cause problems in reality.

1. Is the expected impact upon successful attack severe?
2. Is there a business case for adversaries to attack? Is it clear where a population that represents positive utility for the attacker lies under T?
3. Does the attack scale (is launching multiple instances relatively cheap; low variable costs)?

5.2.2 Control-type counterfactuals

Control-type counterfactuals are of the form "If you'd have had this control (but you didn't), this wouldn't have happened (but it did happen)." Whether we spend on such counterfactuals depends on the question whether the control would indeed have been effective against the (actually materialized) threat.

A particular question concerns the target population. Controls may only be sufficiently cost-effective for a subset. In the medical domain, certain tests or vaccinations are only offered to a sub-population, e.g. based on age or medical condition. In this case, it needs to be clear how to define the sub-population.

1. Is the efficacy of X in doubt when T occurs?
2. Should everyone do X or just some sub-population?
3. Is it clear where a population with positive utility for defenders doing X lies?
4. Do we have a measurement showing $\text{Outcome}(X|C) > \text{Outcome}(\overline{X}|C)$? Can we describe conditions C?
5. Is the claim that X is necessary falsifiable? Can we refute the claim that X doesn't make a difference?
6. How expensive is it to determine if the cost of X is lower than the expected harm reduced?

5.2.3 Combined counterfactuals

Combined counterfactuals are of the form: "If I had attacked you (but I didn't), this control would have prevented problems (but it didn't)." Note that the control didn't prevent problems because the threat didn't occur *and* the control wasn't in place (both are hypothetical). The acceptability of such counterfactuals depends on two questions: (a) whether the threat would indeed cause problems in reality, and (b) whether the control would indeed be effective against the (hypothetical) threat. Therefore, the questions of both abovementioned categories apply. Because threat and control are entangled here, additional questions can be asked on the relation between the cost of implementing the control and the cost of investigating the threat.

1. Is the cost of falsifying the threat condition higher than the cost of implementing X? And for society as a whole? (What if hundreds of organisations would each follow the same line of reasoning?)
2. Is the cost of removing other uncertainties (e.g. probability of attack, severity etc.) large relative to the cost of X and the cost of expected impact.

It is worth differentiating between the uncertainty introduced when referring to an unbounded as opposed to bounded interval of time. The claim that something bad will happen within a certain interval is in principle falsifiable (even though it might be expensive an impractical), but the claim that it will happen is not falsifiable.

5.3 Deciding on acceptability

Whenever a counterfactual security argument is given in a discussion, the following points should be evaluated based on the questions above:

1. the necessity of providing the argument as a counterfactual
2. the validity of the counterfactual construct
3. the credibility of the claims implied by the counterfactual
4. the scope of the argument (which (sub)population is addressed)
5. the costs of acquiring additional information.

First, there should be clear reasons why a counterfactual is presented rather than a more standard form of evidence; it is troublesome if conventional evidence might decide the benefit of the measure but a counterfactual is offered instead. This holds both for the hypothetical occurrence of threats and the hypothetical effect of controls. Second, it is important that the counterfactual comport with what we know of the world; a counterfactual that argues for a measure is suspect if simpler attacks on the same asset do not appear to be exploited. In that case, available evidence suggests that the particular scenario is unlikely to occur, and the effect of the control will therefore be more limited than the counterfactual might suggest. Third, evaluating credibility requires documenting the assumptions in the counterfactual and assessing their likelihood; simple assumptions seem preferable to complex ones, a single assumption is preferable to a counterfactual that involves many; counterfactuals where we have difficulty detailing the precise assumptions are suspect. In particular, hypothetical threats should be clearly distinguished from control requirements (even if threats are expected to occur, this does not imply that controls need to be implemented). Fourth, it should be clear if a counterfactual suggests that a measure should apply to the whole population, or just a portion; a circumstance that

applies to 0.01% of the population probably does not justify forcing all to adopt some measure (see also another paper in this conference on individualized controls: [8]). Finally, we should consider what evidence would argue against the measures that a counterfactual argues for; the more expensive that evidence is to gather, the closer the counterfactual is to unfalsifiable.

We revisit some of the earlier examples in light of these guidelines. There is little difficulty in judging the counterfactual that argues for lifeboats acceptable. It is clear why more standard evidence is not offered; we know that ships sink; the set of assumptions involved is small and easily articulated; we know that the measure should apply to all who travel on the water; historically the measure has survived many potential falsifying events. Password masking has more suspect support. An A/B cohort study might help decide the question, but has not been performed; cheaper and more scalable attacks exist to gather passwords; the assumptions around attacker motivation for such an expensive attack are unclear; it is likely that only a tiny fraction of the population would be valuable enough to attack, and target selection for an attacker would be difficult; the cost of a large cohort study could be considerable.

We cannot provide a fully formalized decision procedure at this stage. Instead, we think that pointing to these criteria should already provide a major advantage to those who are struggling with evaluating security arguments. Further research can refine and improve the questions and criteria based on additional cases, formalization of the constructs, as well as empirical studies of the use of counterfactuals in practice, and their effect in discussions and decision-making processes.

5.4 Related work

In our Dagstuhl seminar report, we have discussed different types of security metrics in terms of type 1 vs. type 2 security metrics [13], without explicitly referring to counterfactuals. Type 1 metrics measure "resistance" (or Vulnerability), whereas type 2 metrics measure security in terms of reduced risk or loss (impact of incidents).

The focus on exclusion or inclusion of the threat environment is related to the work by Böhme [5]. Böhme states that the cost of security maps to benefits of security via an intermediate variable, called "security level". This "security level" is comparable to what we have called a type 1 metric. Böhme does not state explicitly what the role of the threat environment in this mapping is, but it must be the case that the threat environment influences the mapping from security level to the benefits of security.

Connections between security and logic are not new. In particular, logic has been used to reason about access control policies (see e.g. [28, 33, 34]) as well as program security (see e.g. [18, 31]). Even counterfactual logic has been applied in the domain of policies [25], in order to provide change-impact analyses for access control. Here, we focus on logical arguments in empirical security research, which is not so common. In particular, this work is inspired by philosophy of science, notably falsifiability of scientific claims, which has recently been taken up in the security domain (mostly in relation to open research data) [12].

There are some very interesting methodological papers from different fields that address the issue of counterfactuals. In particular, Blundell & Costa Dias discuss how counterfactuals are related to experimental approaches, and how "constructing the counterfactual" is essential when evaluating non-experimental data [4]. Robinson et al. [30] make a similar point, arguing that experimental designs aim at getting as close to observing a counterfactual as possible, meaning a situation "in which *every* factor save one is identical to the former, with the deviating factor being the one of substantive interest to the researcher" (p. 345). These views support our point about the connection between counterfactuals and experiments. However, they do not address the issue of whether manipulating one factor (in our case the threat environment) is informative when making decisions on remedies.

6. CONCLUSIONS AND DISCUSSION

6.1 Conclusions

In this paper, we have shown that counterfactuals are a necessary evil in security reasoning. They are necessary because security is inherently tied to uncertain possibilities of attack by sufficiently motivated attackers. They are evil (or at least dangerous) because they can easily be misused, misinterpreted, or taken out of context in order to support dubious security rhetorics. We have argued that the use of counterfactuals needs to be better understood and better controlled. To this end, we have devised a structured list of questions on the acceptability of specific counterfactuals as security arguments, distinguishing between threat-type and control-type counterfactuals, emphasizing the population assumed in the arguments and the cost of acquiring more information on conditions (falsifying the counterfactual). A summary of counterfactual types is depicted in Table 1.

Compared to the state of the art and standard risk management discourses, our focus on counterfactuals provides three key benefits:

1. A better explanation of the link between security arguments, threats in the real world, and artificially induced threats;
2. A better understanding of the link between security arguments and research methods (such as penetration testing and attack trees);
3. More emphasis on the cost of information gathering (e.g. falsification).

6.2 Discussion

6.2.1 *Counterfactuals and metrics*

As we have shown, the discussion on counterfactuals is inextricably bound to the discussion on security metrics. We can measure the effectiveness of a control by inducing the threat it was meant to protect against, thereby testing the control in an artificial (counterfactual) environment. Or we can measure whether damage/victimization is reduced for those who implement the control (preferably under random assignment, to avoid selection bias).

Also for the analysis methods it matters whether the tests are experimental what-if measurements (e.g. [6]) or counts of actual incidents (e.g. [36]). Quantitative metrics get completely different meanings depending on the type of measurement. In the counterfactual case (induced threat), one can

	Actual defense/control	Hypothetical defense/control
Actual threat	*Actual effect*: I attacked you and this happened	*Control-type counterfactual*: I attacked you and this would have happened if you had this defense/control
Hypothetical threat	*Threat-type counterfactual*: You have this defense/control, and if I had attacked you, this would have happened	*Combined counterfactual*: If I had attacked you, and if you had this defense/control, this would have happened

Table 1: A summary of counterfactual types

for example measure how much time it takes the attacker / penetration tester to succeed. In the non-counterfactual case (actual incidents), one can for example measure how much time it takes until a particular machine is infected. These metrics are completely different, as the former relates to time invested by the attacker, whereas the latter relates to the elapsed real time. In the latter case, the time measured is basically the sum of the time until the threat occurs, plus the time required for the threat to succeed. The time until the threat occurs is zero by definition in the counterfactual situation.

How can one reconcile the two different security argumentation styles? If one knows counterfactual security, then adding a (model of) a threat environment can provide some idea of the likelihood of occurrence of attacks [27]. Whereas the controlled threat of the counterfactuals typically is a single attack (if we would send this particular phishing e-mail...), the threat model needed here must predict how often similar or different phishing e-mails will be sent in practice.

Although a threat model for phishing may be based on frequencies, this does not work for targeted attacks (APTs). In these cases, a model of a strategic attacker should be used as a threat model, which, together with the counterfactuals, provides information on which attacks to expect. Such reasoning is typically based on the expected utility of attack scenarios for the attacker, which may depend on the attacker motivation, resources, and skill. The higher the security level (control strength), the lower the expected utility, and the lower the likelihood of occurrence of attack attempts. A complication here is that many attacks may not be not so targeted that they would not target a different organisation if that organisation has much weaker security.

6.2.2 Burden of proof

As mentioned, counterfactuals are also connected to the burden of proof. The bottom line is that if the possibility that a threat event may occur is used as an argument for the necessity of implementing controls we have no way of bounding defensive spending. The key issue then is where the burden of proof should lie: do we invest by default or not invest by default when it is unclear whether the measure is worth it? Counterfactuals seem to put the burden of proof on those who would not want to implement controls; they would have to prove that the proposed what-if scenario is not credible. However, many of the counterfactuals contain unfalsifiable conditions, i.e. unfalsifiable claims on future threat environments, which makes this burden of proof impossible to satisfy [15]. This is also related to the discussion on the precautionary principe [29].

A particular case in point are the counterfactuals of the type "If you don't have this control, you'd have a problem (if you're attacked)". Although such counterfactuals seem to tell you that you need the control, they don't say anything whatsoever about the effectiveness of the control. You might have exactly the same problem if you do implement it. They are of the type "If you don't eat bananas then you could be robbed." Still, they put the burden of proof for (in)effectiveness on those who do not want to implement it. Note the role of negation in this particular counterfactual type, which could be a topic for further discussion / research.

6.2.3 Counterfactuals as narratives

A particularly interesting feature of the discussion at the workshop was the point that counterfactuals are not just logical statements, but also narratives. As such, they are also *stories* on risks. From this point of view, it is not just the logical validity that matters, but also the narrative power that counterfactuals have. For example, if people do not take seriously a new type of risk (for which there may be valid arguments), counterfactuals can be used to illuminate the possible event and its consequences. In this context, they are used as rhetorical tools for convincing a specific audience, and they are thus part of risk communication. This may also mean that the form of the counterfactual may be tailored for the target audience. On the other hand, the narrative power may also highlight spurious risks, which we have discussed in this paper. If the narrative power of certain counterfactuals is very strong, they may become the focus of an entire discourse, even obscuring other and potentially more important risks. The narrative power of counterfactuals can therefore be both a blessing and a curse.

6.3 Open questions

As this is a new perspective on security arguments, there are plenty of opportunities for further study. In particular, we think there is room for further extending and improving the list of questions as well as the criteria for evaluating counterfactual arguments in security. Another extension would be a more quantitative approach judging the quality of counterfactuals rather than a binary scale of acceptability. Although this complicates the acceptability decision, it leaves room for subtlety in judgement, similar to p-values in statistics, especially when there is a certain level of uncertainty in the claims involved.

Secondly, there are opportunities for investigating the perspective we sketched in this paper to related approaches. A specific possibility for further research is the integration of counterfactuals and specific types of counterfactuals in

argumentation-based approaches to cyber security risk management.

Acknowledgments

The foundations for this paper were laid in the Dagstuhl seminar on Socio-technical Security Metrics [13]. The authors wish to thank Michel van Eeten for helpful suggestions. The research of the second author has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement ICT-318003 (TRE$_S$PASS). This publication reflects only the authors' views and the Union is not liable for any use that may be made of the information contained herein.

7. REFERENCES

[1] http://www.schneier.com/.

[2] F. Arnold, W. Pieters, and M. Stoelinga. Quantitative penetration testing with item response theory. In *Information Assurance and Security (IAS), 2013 9th International Conference on*, pages 49–54, Dec 2013.

[3] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith. Smudge attacks on smartphone touch screens. *WOOT*, 10:1–7, 2010.

[4] R. Blundell and M. Costa Dias. Evaluation methods for non-experimental data. *Fiscal Studies*, 21(4):427–468, 2000.

[5] R. Böhme. Security metrics and security investment models. In *Advances in Information and Computer Security*, pages 10–24. Springer, 2010.

[6] J.-W. H. Bullée, L. Montoya, W. Pieters, M. Junger, and P. H. Hartel. The persuasion and security awareness experiment: reducing the success of social engineering attacks. *Journal of Experimental Criminology*, 11(1):97–115, 2015.

[7] H. Cavusoglu, B. Mishra, and S. Raghunathan. A model for evaluating IT security investments. *Commun. ACM*, 47(7):87–92, 2004.

[8] S. Egelman and E. Peer. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, New York, NY, USA, 2015. ACM.

[9] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner. A survey of mobile malware in the wild. In *Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices*, pages 3–14. ACM, 2011.

[10] P. Finn and M. Jakobsson. Designing ethical phishing experiments. *Technology and Society Magazine, IEEE*, 26(1):46–58, 2007.

[11] D. Florêncio and C. Herley. Where Do All the Attacks Go? *WEIS, 2011, Fairfax.*

[12] D. Gamayunov. Falsifiability of network security research: The good, the bad, and the ugly. In *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering*, TRUST '14, pages 4:1–4:3, New York, NY, USA, 2014. ACM.

[13] D. Gollmann, C. Herley, V. Koenig, W. Pieters, and M. A. Sasse. Socio-Technical Security Metrics (Dagstuhl Seminar 14491). *Dagstuhl Reports*, 4(12):1–28, 2015.

[14] C. Herley. Security, Cybercrime and Scale. *Comm. ACM. Oct. 2014.*

[15] C. Herley. The Unfalsifiability of Security Claims. *Microsoft MSR-TR-2015-72, Sept. 2015.*

[16] D. Ionita, J.-W. Bullee, and R.J. Wieringa. Argumentation-based security requirements elicitation: The next round. In *Evolving Security and Privacy Requirements Engineering (ESPRE), 2014 IEEE 1st Workshop on*, pages 7–12, Aug 2014.

[17] B. Jacobs and W. Pieters. Electronic voting in the Netherlands: From early adoption to early abolishment. In A. Aldini, G. Barthe, and R. Gorrieri, editors, *Foundations of Security Analysis and Design V*, volume 5705 of *Lecture Notes in Computer Science*, pages 121–144. Springer Berlin Heidelberg, 2009.

[18] B. Jacobs, W. Pieters, and M. Warnier. Statically checking confidentiality via dynamic labels. In *WITS '05: Proceedings of the 2005 workshop on Issues in the theory of security*, pages 50–56, New York, NY, USA, 2005. ACM Press.

[19] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, et al. Experimental security analysis of a modern automobile. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 447–462. IEEE, 2010.

[20] B. Lampson. Usable security: how to get it. *Communications of the ACM*, 52(11):25–27, 2009.

[21] S. Mauw and M. Oostdijk. Foundations of attack trees. In D. Won and S. Kim, editors, *Proc. 8th Annual International Conference on Information Security and Cryptology, ICISC'05*, volume 3935 of *Lecture Notes in Computer Science*, pages 186–198. Springer, 2006.

[22] M. Backes, M. Duermuth, and D. Unruh. Compromising Reflections - or - How to Read LCD Monitors Around the Corner. *IEEE Symposium on Security and Privacy*, 2008.

[23] M. Naor. On cryptographic assumptions and challenges. In *Advances in Cryptology-CRYPTO 2003*, pages 96–109. Springer, 2003.

[24] M. E. Paté-Cornell. Fault trees vs. event trees in reliability analysis. *Risk Analysis*, 4(3):177–186, 1984.

[25] M. Peralta, S. Mukhopadhyay, and R. Bharadwaj. Counterfactually reasoning about security. In *Proceedings of the 4th International Conference on Security of Information and Networks*, SIN '11, pages 223–226, New York, NY, USA, 2011. ACM.

[26] C. P. Pfleeger and S. L. Pfleeger. *Security in computing*. Prentice Hall Professional, 2003.

[27] W. Pieters and M. Davarynejad. Calculating adversarial risk from attack trees: Control strength and probabilistic attackers. In *3rd International Workshop on Quantitative Aspects in Security Assurance (QASA)*, Lecture Notes in Computer Science. Springer, 2014.

[28] W. Pieters, T. Dimkov, and D. Pavlovic. Security policy alignment: A formal approach. *Systems Journal, IEEE*, 7(2):275–287, 2013.

[29] W. Pieters and A. van Cleeff. The precautionary principle in a world of digital dependencies. *IEEE Computer*, 42(6):50–56, 2009.

[30] G. Robinson, J. E. McNulty, and J. S. Krasno. Observing the counterfactual? the search for political experiments in nature. *Political Analysis*, 17(4):341–357, 2009.

[31] A. Sabelfeld and A.C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.

[32] B. Schneier. Attack trees: Modeling security threats. *Dr. Dobb's journal*, 24(12):21–29, 1999.

[33] M. Sloman and E. Lupu. Security and management policy specification. *Network, IEEE*, 16(2):10 –19, 2002.

[34] B. Solhaug and K. Stølen. Preservation of policy adherence under refinement. *Int J Software Informatics*, 5(1-2):139–157, 2011.

[35] The Open Group. Risk taxonomy. Technical Report C081, The Open Group, 2009.

[36] M. J. G. Van Eeten, J. Bauer, H. Asghari, and S. Tabatabaie. The role of internet service providers in botnet mitigation: An empirical analysis based on spam data. OECD STI Working Paper 2010/5, Paris: OECD, 2010.