SoK: Science, Security, and the Elusive Goal of Security as a Scientific Pursuit

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

P.C. van Oorschot
Carleton University, Ottawa, ON, Canada
paulv@scs.carleton.ca

***Abstract*—The past ten years has seen increasing calls to make security research more "scientific". On the surface, most agree that this is desirable, given universal recognition of "science" as a positive force. However, we find that there is little clarity on what "scientific" means in the context of computer security research, or consensus on what a "Science of Security" should look like. We selectively review work in the history and philosophy of science and more recent work under the label "Science of Security". We explore what has been done under the theme of relating science and security, put this in context with historical science, and offer observations and insights we hope may motivate further exploration and guidance. Among our findings are that practices on which the rest of science has reached consensus appear little used or recognized in security, and a pattern of methodological errors continues unaddressed.**[1]

***Index Terms*—computer security research; science of security; history of science; philosophy of science; connections between research and observable world.**

I. Introduction and Overview

Security is often said to have unique challenges. Progress can be harder to measure than in areas where, e.g., performance metrics or capabilities point to visible steady improvement. Supposedly unique factors, such as the presence of active adversaries, complicate matters. Some even describe the field in pessimistic terms. Multics warriors remind the young that many of today's problems were much better addressed forty years ago [1]. Shamir, in accepting the 2002 Turing award, described non-crypto security as "a mess." Schell, in 2001, described the field as being filled with "pseudo-science and flying pigs" [2].

Perhaps in response to these negative views, over the last decade there has been an effort in parts of the community to develop a "Science of Security" (SoS). In this paper we review both work in the history/philosophy of science and, recently, under this SoS banner. We wish to distinguish at the outset between these two strands. The first is an exploration of the techniques that the consensus from other fields suggest are important to pursuing any problem scientifically. The second is the activity and body of work that has resulted from external promotion of an agenda by the name "Science of Security". It is not our objective to argue directly for, or against, the work under this label. Rather, given the effort by several governments to promote and fund an agenda under this name, we explore what has been done and how it has been pursued. This leads us to consider the program (and other security research) in the light of consensus views of science and scientific methods. We find that aspects from the philosophy of science on which most other communities have reached consensus appear surprisingly little used in security, including in work done under the SoS label. For example, we do not find that that work better adheres to scientific principles than other security research in any readily identifiable way.

We identify several opportunities that may help drive security research forward in a more scientific fashion, and on this we are cautiously optimistic. While we see great benefit to this, we also do not wish to argue that all of security must be done on rigidly scientific principles. A significant component of security is engineering; this shares with science the regular contact with, and feedback from, observation, despite not having as clearly articulated a definition or methods.

Section II selectively reviews literature on the history and philosophy of science, with particular emphasis on three things: 1) methodologies and positions on which practicing scientists and philosophers of science have largely reached consensus; 2) aspects highlighting opportunities to eliminate confusion in security research; and 3) contributions pointing to where security research might be made "more scientific". Section III selectively reviews literature relating "science" and "security", for examples of viewpoints within the community, for context in later discussion, and as supporting evidence for arguments; an exhaustive review of all security literature attempting to determine which papers use scientific methods in security research is not a goal. Section IV highlights areas where the security community has failed to adopt accepted lessons from the science literature. Section V provides insights and offers observations and constructive suggestions. Section VI concludes.

II. History/Philosophy of Science

This section highlights aspects from the history and philosophy of science most relevant to security research. Our goal here is not an encyclopedic review of science literature; accessible summaries are available in introductory books by, e.g., Chalmers [3] and Godfrey-Smith [4]. We ask patience of readers who might question the relevance of this material to security; Sections IV and V show that neglect of these lessons is at the root of several significant problems.

[1] March 22, 2017. Author's copy for personal, non-commercial use. This paper will appear in: Proc. 2017 IEEE Symposium on Security and Privacy.

A. Separation of Deductive and Inductive Statements

Probably the most significant settled point in the philosophy of science is that inductive and deductive statements constitute different types of knowledge claims. That is, we draw conclusions about the empirical world using observations and inferences from those observations, and these are fundamentally different from mathematical or deductive statements derived from axioms. For example, after many observations we may infer rules which account not just for the things observed, but things not yet observed (e.g., "all swans are white"). These can always turn out to be wrong, if a future observation violates a rule we have inferred (e.g., we observe a black swan). Deduction, by contrast, produces statements that follow with certainty from a self-consistent set of axioms. An example is Euclidean geometry; e.g., Pythagoras' theorem follows from the axioms, and there is no possibility of observing anything that violates it. (Section II-C examines the claim that Geometry actually describes the world.)

The importance of this distinction has long been known. Versions date back to Plato, who distinguished the messy, physical realm of things observed from the perfect non-physical world of "forms." Western classical thought was heavily influenced by the view that universal truths were found only by reasoning about forms (vs. the ephemeral world of observable things); hence, perhaps, the emphasis on Logic and Geometry.

The separateness of these real and ideal realms is known by different names. Kant's very thorough treatment was influential [5]; he calls statements whose truth is independent of experience *a priori* and those that depend on observation *a posteriori*. The modern description relies heavily on the work of the *logical positivists*, an influential group active in Vienna in the early 1900's; they called it the *analytic-synthetic distinction*. An *analytic proposition* is one whose truth follows with certainty if the premises are true; such propositions are associated with *deductive* statements and include logical truths or syllogisms. In contrast, a *synthetic proposition* is one whose truth depends on the relationship between its meaning and the real world; these are associated with *inductive* statements, their truth or falsehood dependent on empirical observations [6].

While deductive statements are certain consequences of the premises, inductive statements are always subject to error. (Note: despite misleading terminology, the mathematical technique of proof by induction is deduction.) Considerable philosophy of science literature examines the question of when we can rely on an inductive statement. Hume's "problem of induction" [7] is that the logical basis for believing inductive claims is weak in comparison with the certainty of deductive ones. No amount of corroborating evidence can establish the truth of a generalization. That all the swans we have seen were white is no guarantee against encountering a black one. That induction has proved a reliable guide to knowledge before is itself an inductive argument, and thus circular.

Departures from the classical view accelerated with the rise of empirical approaches to knowledge discovery driven by scientists such as Galileo. His discovery of the moons of Jupiter posed a major challenge to Aristotle's assertion that all heavenly bodies orbited the earth. Bacon [8] formalized an inductive method of generalizing from observations—his method of observation, generalization and correction is acknowledged as an early articulation of what many historically view as the basic scientific method (see Section II-E). This stood in contrast to the classical approach of deduction from things that were assumed true (e.g., that objects fall at a speed proportional to their mass as asserted by Aristotle).

While Plato had argued that what we could observe was only an illusory shadow of the perfect world of forms, Mill argued essentially the reverse: induction is our only path to understanding the world since, on its own, deduction is incapable of helping us discover anything about the world [9]:

> *But this is, in fact, to say that nothing ever was, or can be, proved by syllogism, which was not known, or assumed to be known, before.*

Ayer, a logical positivist, summarizes [6, p.57]:

> *the characteristic mark of a purely logical inquiry is that it is concerned with the formal consequences of our definitions and not with questions of empirical fact.*

This is now well-established, and recognized by scientists from fields as diverse as Physics [10], [11], Biology [12], and Economics [13]. Serious scientists do not claim certainty for their statements about the world, and do not claim to deduce facts about the world that weren't implicit in the assumptions. To quote Medawar [12]:

> *Deduction in itself is quite powerless as a method of scientific discovery—and for this very simple reason: the process of deduction as such only uncovers, brings out into the open, makes explicit, information that is already present in the axioms or premises from which the process of deduction began.*

At the risk of laboring the point, recall Einstein:

> *As far as the laws of Mathematics refer to reality they are not certain, and as far as they are certain they do not refer to reality.*

Deduction does not allow discovery of *new* facts about the world. Axioms and definitions are not real-world facts, so deduction starting there can say nothing at all about the world. On the other hand, deduction that begins with assumptions or inductive inferences can explore their real-world implications. Thus, deductions that start from Newton's laws allow conclusions about real-world observations, but deductions from Euclid's postulates do not.

B. Falsification as Demarcation Criterion

Thus, pure deduction offers no route to reliable knowledge about the world. Induction, on the other hand, does allow general statements about the world, but we can never be sure that they are true. This allows room for statements that are sensible and well-grounded, but also ones that have little basis. Popper sought an answer to the question of demarcation, i.e., a clear criterion which separates scientific from non-scientific theories. The answer clearly could not be that scientific statements are true and non-scientific ones false; he knew [14] "that science often errs, and that pseudoscience may happen to stumble on the truth." For example, we now know that both Thompson's "plum-pudding" model of the atom and Pauling's triple helix model of DNA are incorrect, but it would seem harsh to describe them as unscientific for being inconsistent with observations that weren't available at the time. A simplified statement of Popper's criterion is that scientific theories should be falsifiable [14]:

> *A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice.*

In Popper's view, to count as scientific a theory must "stick its neck out" and make predictions or statements that run at least some risk of being contradicted by empirical observation. A theory not running such a risk is compatible with every possible set of observations. Popper suggested that Marxist theory and Freudian analysis were non-scientific since they were elastic enough that no observation ever seemed to adherents incompatible with the predictions; Einstein's Relativity, by contrast, even though it assaulted basic intuitions about time and space, made testable predictions.

This criterion, eliminating things that "can't be brought into contact with observation", places a number of theories firmly in the non-scientific camp. We easily classify claims about the existence of UFO's, the Loch Ness monster, and paranormal phenomena as unfalsifiable; failure to stake their accuracy on any test they might actually fail rules them unscientific in Popper's view. Religious and metaphysical claims are also separated from Science by this criterion; so too, more surprisingly, is Mathematics (see Section II-C).

In addition to insisting on falsifiability, it is often said that Science progresses by finding errors. According to Popper [14] "in finding our conjecture to be false we shall have learnt much about the truth." Thus, scientists emphasize efforts at refutation rather than confirmation. This has the consequence that it forces theories to be precise [3, pp. 63-64]; the less precise a claim the easier it is to corroborate and the harder to falsify (as it is consistent with more things). The claim "no true scotsman puts sugar in his porridge" is corroborated by every scotsman who foregoes sugar, but impossible to refute since what counts as "true" is left open. Stating the evidence that would falsify a claim thus acts as a forcing function to clarify vague claims and implicit assumptions.

It would be simplistic to regard falsification as a final answer to question of what counts as science. It can be difficult to agree what precisely counts as a falsification, so the criterion isn't as clear as we might like. For example, Popper at one point wrote [15, p.151] "Darwinism is not a testable scientific theory but a metaphysical research programme", only to reverse himself later [16, p.345]:

> *I have changed my mind about the testability and the logical status of the theory of natural selection; and I am glad to have an opportunity to make a recantation.*

Thus, while some statements are clearly unfalsifiable, many gray areas are open to dispute. Problems with falsification are discussed further in Appendix A.

C. Relation between Mathematics and Science

That falsification as a criterion classifies Astrology and Homeopathy as non-science seems correct. That it also classifies Mathematics (and all deductive statements) as non-science is more jarring. This can seem strange. Some of the greatest achievements of Science, as embodied in the work of Newton, Maxwell, and Einstein seem inseparable from their mathematical expression. Since a great deal of Computer Science is heavily mathematical (see Section II-C2), it is worth seeking clarity on this point.

Since mathematical statements cannot be contradicted by any real-world observation, they are compatible with every observation, and thus can make no claims and can offer no guarantees about anything in the real-world. Ayer warns about the philosopher "who posits certain first principles, and then offers them with their consequences as a complete picture of reality" [6, p.46]. Euclidean geometry represents a case where it is very tempting to regard a purely deductive system as describing the world around us. Ayer points out that even this view is insupportable, since, e.g., with different replacements for Euclid's parallel postulate, we end up with not one, but many non-Euclidean geometries each with a plausible claim to describe reality [6, p.82]:

> *Whether a geometry can be applied to the actual physical world or not, is an empirical question which falls outside the scope of the geometry itself. There is no sense, therefore, in asking which of the various geometries known to us are false and which are true. In so far as they are all free from contradiction, they are all true.*

Continuing, Ayer elaborates that a deductive system cannot guarantee statements about the world, including statements about how well it resembles the world:

> *But the proposition which states that a certain application of a geometry is possible is not itself a proposition of that geometry. All that the geometry itself tells us is that if anything can be brought under the definitions, it will also satisfy the theorems.*

This is a limitation of any mathematical model: how well a model matches reality can only be tested empirically. For example, Shannon cautioned that whether and how well Information Theory applied to any problem was a matter that could only be decided empirically [17]: "If, for example, the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations."

Thus, statements about the real-world derived from deductive systems are subject to some uncertainty about how well the latter matches the former. We can use the fact that Newton's laws appear to describe reality very accurately to generate predictions, but those predictions inherit all of the uncertainty about the assumed fit of the model (recall the quotations from, e.g., Einstein in Section II-A).

*1) **Accuracy of predictions over realism of assumptions**:* Mathematical models have proved enormously useful in various branches of Science. It is worth emphasizing that a mathematical model is judged on the accuracy of its predictions rather than the perceived reasonableness of its assumptions. There is no confusion on this point when checking predictions against measurements is easy: when there's a conflict between the model and the data, the data wins (see, e.g., remarks by Feynman and others in Section II-D). However, in disciplines where direct measurement and controlled experiments are hard (e.g., Economics and Security) it can be tempting to seek other forms of validation for a mathematical model. Friedman reminds us that whether assumptions appear reasonable does not represent an alternative, or additional, route to deciding the reliability of a theory (since this is a subjective assessment) [13]: "What is the criterion by which to judge whether a particular departure from realism is or is not acceptable?" He observes that Newton predicts that *in a vacuum* the speed of a body falling toward the earth will be $g \cdot t$. Of course, the air pressure is not zero; so is the assumption of a vacuum reasonable? At the surface of the earth it would appear the answer is "yes" for a brick, but "no" for a feather. Some very successful theories are based on assumptions that appear hard to justify; e.g., with radius 6.96×10^8 metres, does the sun qualify as a point mass? Conversely, what appear to be reasonable assumptions can turn out to be very wrong.

Of course, this does not imply that reasonable assumptions are no better than unreasonable ones. It simply serves to remind us that models stand or fall on the accuracy of their predictions. Just because measurements to test predictions are hard to conduct does not mean a model can be assumed to pass or be exempted from this test. That an assumption is reasonable, or an approximation "good enough", is unfalsifiable unless an explicit test is committed to.

*2) **Programs as predicates, now with real-world inputs**:* The origins of Computer Science and Computer Security are mathematical. WWII ciphers and code-breaking are justly revered. Without the accomplishments of modern cryptography, it is doubtful the World-Wide-Web would have attained such influence. This has contributed to the strongly mathematical flavor of our discipline. In 1984, Hoare describes a computer program as [18] "a logical predicate describing all of its permitted behaviors"; and further, that the final design meets the original requirements "can be mathematically proved before starting the implementation of its components." This view of programs as implementing mathematical predicates, is unsurprising given the importance of the development of algorithms. For an algorithm that involves sorting, searching etc., it is important that its behavior is well-understood under all possible inputs.

This may be possible in a closed-world setting. However, the range of inputs to consider is much more complex with today's applications. Programs depend not just on well-structured data, but human (possibly adversarial) input and program behavior across distributed networks. The successes of many machine learning applications such as face and speech recognition, and self-driving cars, rely heavily on the ability to train on unconstrained real-world data sets. As Pavlovic notes [19], viewing programs as predicates is no longer appropriate in these settings. In preliminary attempts to address this, recent work is exploring protocol analysis incorporating human actions through *ceremony analysis* of Ellison [20] (see also [21], [22]). We note that a decade after the comments above Hoare appears to have had a change of heart (see Section III-A) [23].

D. Viewpoints of Major Scientists

Many philosophers of science regard Popper's analysis as incomplete, e.g., due to the complications outlined in Appendix A. However, the centrality of falsification in deciding what is and is not science is supported by its hold on the minds of practicing scientists. It seems safe to say most scientists' understanding of the philosophy of science terminates with Popper, possibly excepting work on social structure largely triggered by Kuhn [24].

Front-rank scientists of the 20th century who wrote or spoke about the process of science seem to stick largely to the view that falsification defines the boundary. Bohr, who probably wrestled with philosophical issues more than any leading modern physicist, identified with the logical positivists [25], and regarded questions that could not be tested as non-scientific. His dictum [25] "It is wrong to think that the task of physics is to find out how Nature is. Physics concerns what we can say about Nature" appears an explicit recognition that pondering questions we can't verify or falsify is pointless.

Feynman's writings provide many comments on his view of the scientific method. In a 1964 lecture at Cornell, Feynman's description of the scientific method is almost explicitly a summary of Popper's *Conjectures and Refutations* [14] and closely resembles Section II-E's hypothetico-deductive model:

> *In general, we look for a new law by the following process. First, we guess it, no, don't laugh, that's really true. Then we compute the consequences of the guess, to see what, if this is right, if this law we guess is right, to see what it would imply and then we compare the computation results to nature, or we say compare to experiment or experience, compare it directly with observations to see if it works. If it disagrees with experiment, it's wrong. In that simple statement is the key to science.*

An influential paper by Platt [26], advocating what he calls *strong inference*, largely echoes Popper in the desire to seek ways to rule things out: "any conclusion that is not an exclusion is insecure." Platt points out that even scientific fields often fall short and fail to progress rapidly. He advocates that when offering an observation we explicitly state what hypothesis it refutes, and when offering a hypothesis we explicitly state what observation would refute it.

Several remarks from major scientists emphasize that a scientific claim seeks out rather than shrinks from tests that might prove it wrong. Pauli's famous description of a paper as being "not even wrong" is usually taken as a criticism of something that is not falsifiable or makes no testable prediction. Similarly, Feyman's famous quip [11] "You can't prove a vague theory wrong" appears to reiterate that he largely agreed with Popper on the question of falsification and the importance of finding errors.

While Darwin lived and worked before the emphasis on falsification as a criterion, what he wrote [27] on his approach to Science—that he proceeded "on true Baconian principles"—fits well with a modern understanding. He also mentioned the importance of noting ideas that conflicted with his theory for fear that he might later forget them [28], apparently endorsing the idea that Science advances by finding errors.

Falsification is now so firmly established across various branches of Science that perceived deviations meet with harsh rebukes. Evolutionary biologist Ayala writes [27]:

> *A hypothesis is scientific only if it is consistent with some but not other possible states of affairs not yet observed, so that it is subject to the possibility of falsification by reference to experience.*

Ellis and Silk suggest that those not adhering to it threaten the integrity of Physics [10]:

> *In our view, the issue boils down to clarifying one question: what potential observational or experimental evidence is there that would persuade you that the theory is wrong and lead you to abandoning it? If there is none, it is not a scientific theory.*

E. Hypothetico-deductive Model (basic scientific method)

Minimally, to count as scientific, we expect a theory to have the following properties:

- *Consistency:* claims are consistent with other claims and available observations. Inconvenient observations are not discarded.
- *Falsifiability* (see above): we can describe the evidence that would prove claims wrong. Without this we are not self-correcting [14].
- *Predictive power and progress:* models and theories should facilitate accurate predictions, and the set of observations that can be accurately predicted should generally increase over time. We should not be asking all of the same questions year after year.

By this, an inconsistent or unfalsifiable theory, providing neither new understandings nor accurate predictions, or a field that doesn't progress, is unscientific. We separate these properties from the means often used to achieve them, e.g., openness and data-sharing, peer review, reproducible experiments, etc.

Although declaring a one-size-fits-all-problems recipe is overly simple (and less pronounced in newer science textbooks [4, pp.6-7]), emphasis on consistency and falsification led to what remains a popular perception of ("a" or "the") *scientific method*. The idea is to attempt to generalize, making falsifiable statements that are consistent with what we have already observed, but predict also things not yet observed. Then seek new observations, especially those expected to present severe tests of predictions (rather than those expected to corroborate them). Often called the *hypothetico-deductive* model, the summary is:

1) Form hypotheses from what is observed.
2) Formulate falsifiable predictions from those hypotheses.
3) If new observations agree with the predictions, an hypothesis is supported (but not proved); if they disagree, it is rejected.

Some variant of this is likely the closest available to a consensus on how Science works (see Section II-D), despite naive falsification having known issues (see Appendix A).

This model is simply a method delivering the desired properties above—inconsistencies are rooted out, all claims are considered fallible, knowledge improves iteratively. Note that this process of iteratively eliminating possibilities that conflict with observations is the essence of differential diagnosis in medicine, sensible approaches to car repair, and the investigative method Sherlock Holmes recommends to Watson: "Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth."

F. Sciences of the Artificial (Research on Human Artifacts)

Here we consider how Security research relates to traditional scientific fields, and the generality and longevity of results.

Natural science is dominated by the physical sciences (Physics, Chemistry, Astronomy, and newer Earth Science), and life sciences headlined under Biology. These are distinct from tools of logic, formal reasoning and mathematics, applied sciences (e.g., Engineering), interdisciplinary sciences (e.g., Cognitive Science), and social sciences including Psychology. The place of much younger Computer Science (II-C2) is less clear, with debate on its status as a science [29]. Computer Security is interdisciplinary, using Computer Science, Engineering, Mathematics, and Cognitive Science among others.

Differences between fields and methodologies have led to historical tensions between the pure and applied sciences. Of particular relevance, computer hardware and software are human-made artifacts, part of what Simon called *sciences of the artificial*—activities that involve making artifacts with desired properties. His *design sciences* are those in which man-made objects play a central role—not only Engineering, but also, e.g., Architecture, Education, Medicine, all fields involving design in the sense of "intellectual activity that produces material artifacts". Simon had strong views on pure-applied tensions, and the artificial sciences [30, pp.111-112]:

> *the natural sciences almost drove the sciences of the artificial from professional school curricula, a development that peaked about two or three decades after the Second World War. Engineering schools gradually became schools of physics and mathematics; medical schools became schools of biological sciences; business schools beame schools of finite mathematics ... academic respectability calls for*

> *subject matter that is intellectually tough, analytic, formalizable, and teachable. In the past much, if not most of what we knew about design and about the artificial sciences was intellectually soft, intuitive, informal, and cook-booky.*

What began as intellectually soft ("cook-booky") is now Computer Science—whether science or not, it is now much more rigorous and mature. Security is much less both of these; perhaps we should not be surprised if it is not yet scientific.

Much of Security research is directly dependent on human-made artifacts; e.g., results often depend directly on specific software. This raises the danger that results positioned as general, beyond such artifacts, may be done so incorrectly (naively or optimistically), as generalization is typically inductive. Independent of this, many results in Security may fail to enjoy the long-term relevance of fundamental results in Physics or Chemistry; the atomic structure of elements has changed less in the past ten million years than computer ecosystems and adversaries in the past ten. Thus both the generality and longevity of results impact the type of scientific results we might expect in Security.

G. Pasteur's Quadrant (uniting basic and applied research)

Tensions between basic and applied research did not end with Simon (II-F). Basic research typically has no specific practical goals; applied research addresses specific needs or uses. Stokes [31] rejected this either-or choice and associated *linear model* whereby fundamental research necessarily precedes consideration of practical uses. Observing that science history contradicts the idea of inherent conflict between basic and applied research, Stokes advocated for intersecting fundamental and applied research, as *use-inspired basic research*.

His model of *Pasteur's Quadrant* replaces the linear model by one with a two-by-two grid with Yes/No entries, whose rows ask the question "Quest for fundamental understanding?", and columns ask "Considerations of use?" Yes-No is Bohr's quadrant (his modeling of atomic structure exemplifies basic research); No-Yes is Edison's quadrant (his pursuit of commercial electric lighting exemplifies applied researchers uninterested in basic science implications); Yes-Yes is Pasteur's quadrant (his many contributions combine pursuit of basic knowledge and target use). Stokes writes:

> *as Pasteur's scientific studies became progressively more fundamental, problems he chose and lines of inquiry pursued become progressively more applied.*

Stokes noted that much of modern medical research involves use-inspired basic science spanning these three quadrants—for example, the success of cardiac surgery required a long list of technical innovations across these three.

Regarding Simon's comments (Section II-F) on tensions between pure and applied sciences, Stokes noted that historical bias of researchers against practical use dates back to Greek elites and philosophers, who sought "general forms or ideals rather than" solutions through objects in the physical world; they favored pure inquiry, leaving manual labor (and practical arts) to lower economic classes. Little knowledge was applied for the benefit of society other than doctors practicing medicine. Stokes relates that when Greek science eventually came to western Europe, its "view of superiority of pure science" and corresponding gap between laborers and philosophers was removed by Bacon and contemporaries who believed in both science and manual service. European artisans and pure scientists thereafter jointly improved technology, where previously improving technology was for laborers. This merging of science and technology circa 19C involved many including Kelvin and Maxwell; instances where improved technology led to advances in Science (rather than vice-versa), such as the development of telescopic lenses aiding Astronomy, is given as evidence of historical benefits of combining basic and applied research, neither always occurring first.

III. Science of Security

While many open questions remain in the Philosophy of Science, much is also settled. Far less is settled in discussing Science of Security; what emerges from review of the security literature below is, in many places, an absence of consensus; we return later to consider if this signals an immature science.

A. Science of Security: Early Search and Misunderstandings

Basic principles for Computer Security were set out in now-classic papers, such as Saltzer-Schroeder 1975 [32]. Security research in the 1980s was heavily influenced by the U.S. government *Orange Book* [33] and highly influential *Multics* operating system dating from the 1960s (see the Karger-Schell 1974-2002 bookend papers [1], [34]); at major conferences, Multics seniors still remind the young that today's problems are not only 40 years old, but were better addressed by Multics.

The desire to pursue security research more scientifically is as old as the field itself, and warnings that we have been falling short are just as old and recur often. Already in 1987, McLean [35] decries the poor state of security research with respect to implicit assumptions:

> *Hence, we have developed an environment where our documented foundations are inadequate, yet shielded from adversity by appeals to implicit assumptions "which everybody knows about" (even if people disagree on what these assumptions are!) ... Such is the path to neither science nor security.*

This was part (see also [36]) of a major community controversy suggesting serious flaws in the foundational Bell-LaPadula formal model for confidentiality and challenging the traditional definition of security itself (and a secure system based on the notion of secure states); the flaws were disputed by Bell [37] among others (his Bell-LaPadula retrospective [38] does not cite McLean). McLean notes [39] this divided the community with one side seeing Bell-LaPadula "primarily as a research tool developed to explore the properties of one possible explication of security" while others viewed it as "correctly capturing our informal concept of security"; and that this split "resembled Thomas Kuhn's descriptions of paradigm shifts in science where two communities fail to understand each other because of differing assumptions" [24].

Community reflection included Good's 1987 essay [40], which the *Computer Security Foundations* workshop credits as an inspiration for its foundation. He notes, on the challenge of using system models, that properties proven "may or may not hold for the real system depending on the accuracy of the model", and that for high-level abstractions,

> *What is important is that the abstractions are done in such a way so that when we prove some property about the abstraction, then that property is true of the real, running system.*

(Compare with Ayer's comments on Geometry in Section II-C.) He calls for greater clarity and precision in definitions; recognition of the limitations of models; and recognition that the end goal is to build secure systems (which he asserts requires clear and precisely stated security requirements which are then used as the criteria by which to judge success). Good paints the way forward as formal verification techniques, carefully defined as "the use of rigorous mathematical reasoning and logic in the system engineering process to produce real systems that are proved to meet their requirements." A more-or-less opposite view by DeMillo et al. [41] in 1979 cautions that program verification is bound to fail, and that

> *formal verifications of programs, no matter how obtained, will not play the same key role in the development of computer science and software engineering as proofs do in mathematics.*

They warn of terminology leading to misunderstandings (on formal methods: "it would help if they did not call their verifications 'proofs'"), and note the need to distinguish the deductive and empirical realms:

> *Scientists should not confuse mathematical models with reality—and verification is nothing but a model of believability.*

We note that these arguments and observations have recurred over time and across fields (specifically related to automated verification and automated theorem-proving, see further discussion in Appendix C). In 1989, Schaefer [42] also notes limitations in applying formal methods to trusted systems:

> *the formal analysis process is largely one of working within the rules of a formal system that achieves its results by manipulating uninterpreted symbols. Extensive review of the meanings and implications of the tautologies produced by this process is needed before a supportable conclusion can be drawn on the* relevance *and* applicability *of such results. Equally important, the scope of the formal findings as well as the assumptions on which they are based need to be understood.*

He also comments on the "potentially harmful side-effect of formally designed systems" and warns:

> *If too much faith is placed on the value of using formal methods, and rigorous additional security testing is not borne to bear on candidate systems, it is possible that the only security to be derived from the resultant systems will be a false sense of security.*

Schaefer's 1993 position paper notes the concern [43]:

> *that too much attention was being paid to the manipulation of symbols and too little attention was being focused on the real requirements and actualised properties of the implementation and its platform*

and also notes a number of problems, many still open 24 years later—an interesting signal, if progress defines a science. Hoare in 1996 [23] appeared to have had a change of heart about the importance of formal approaches that he claimed twelve years earlier (cf. II-C2) [18]: "It has turned out that the world just does not suffer significantly from the kind of problem that our research was originally intended to solve."

Kuhn's 1962 book *Structure* [24] is also cited (cf. [44]) for guidance for "disciplines whose fundamental paradigms are in trouble" by Blakley [45], who in 1996 notes shortcomings of the traditional perimeter security model in the face of evolving Internet architectures. Blakley summed up the state of affairs in Computer Security as dismal: "The same exposures keep recurring; we make no practically useful progress on the hard problems". Schell's 2001 essay [2] separates research successes from their use in practice:

> *The state of the science of information security is quite rich with solutions and tools that represent the accumulated knowledge from research over more than 30 years. The state of our assimilation of that knowledge by information security practitioners and understanding of the existing Science is very poor.*

B. Science of Security: Recent Efforts

Here we selectively review research since 2008 under the label "Science of Security"; our goal is not an encyclopedic review per se, but to provide context for later observations.

Over the past 10 years, calls for stronger cybersecurity foundations have increasingly sought "more science", with visible examples including special panels, government-facilitated workshops, funded research programs, and special issues of magazines. A review of these materials finds that despite all this, there remains very little discussion and no consensus definition, of what Science of Security actually entails.

From various organizational and individual calls for more science in security, it is unclear if the general intent is to rule certain types of work in or out, or to emphasize some directions and/or methodologies at the expense of others. Geer, one of the NSA Science of Security prize judges, notes [44]: "our views of what constitutes a, or the, Science of Security vary rather a lot." He continues:

> *Some of us would prioritize purpose ... Some of us view aspects of methodology as paramount, especially reproducibility and the clarity of communication on which it depends. Some of us are ever on the lookout for what a physicist would call a unifying field theory. Some of us insist on the*

> *classic process of hypothesis generation followed by designed experiments. We vary, and I take that to be a vote of sorts on whether cybersecurity is yet a science.*

Evans and Stolfo note [46]: "We're a long way from establishing a science of security comparable to the traditional physical sciences, and even from knowing whether such a goal is even possible." Peisert and Bishop observe [47] "researchers frequently fail to follow the scientific method to support the claims they make in scientific, peer-reviewed papers." The desire to do security more scientifically fits into a larger picture of frustration voiced by experts—e.g., in accepting the Turing award in 2002, Shamir made a set of 10-year predictions that included "the non-crypto part of security will remain a mess."

JASON report of 2010: An 88-page report resulted from a government-solicited exploration of how security research might benefit from greater emphasis on Science [48, abstract]:

> *JASON was requested by the DoD to examine the theory and practice of cyber-security, and evaluate whether there are underlying fundamental principles that would make it possible to adopt a more scientific approach, identify what is needed in creating a science of cyber-security, and recommend specific ways in which scientific methods can be applied...*

Among its recommendations and views was [48, p.77]:

> *The science seems under-developed in reporting experimental results, and consequently in the ability to use them. The research community does not seem to have developed a generally accepted way of reporting empirical studies so that people could reproduce the work.*

The report's two dominating topics were: 1) the immune system analogy; and 2) model checking and related formal methods. The former stressed the need for adaptive responses, using a mixture of sensing modalities for detecting threats rather than relying on one, and the importance of controlled experiments. The latter received strong endorsements tempered by reminders of unsolved issues, aspects still "fraught with difficulties" [48, p.53], and acknowledgements of the research/real world gap: "missing is a more direct connection to the day-to-day practice of programming...there is a need to translate the developments into tools that can be used by those writing software" [48, p.55]. The report was informed by 18 expert presentations, ranging from theorical/formal topics to the experimental/applied. Little mentioned is web or systems security, malware, software management, or human factors.

The significant government-led (especially U.S.) "Science of Security" effort has claimed early success [49], arguably on little evidence. To support this view, or alternately, allow readers to form their own view, Appendix D summarizes literature related to major such initiatives.

C. Claims of What We Need More or Less Of

Several authors explicitly tackle what a Science of Security would entail—with many views and ideas, but less consensus.

*1) **Formal approaches**:* That we need greater reliance on formal approaches is concluded by Good [40], the JASON report [48], and the 2011 National Science and Technology Council (NSTC) Strategic R&D plan [50] (see Appendix D). Observing that current systems "put individuals, commercial enterprises, the public sector, and our military at risk", Schneider [51] suggests "The obvious alternative is to build systems whose security follows from first principles", adding "The field of cryptography comes close to exemplifying the kind of science base we seek." Krawczyk writes [52]:

> *By its very nature, there is no (and cannot be) empirical evidence for the security of a design. Indeed, no concrete measurements or simulations can show that attacks against a cryptographic scheme are not feasible. The only way to do so is to develop a formal mathematical model and language in which to reason about such schemes.*

*2) **Empiricism and data collection**:* Another camp stresses data collection and empirical work. As noted earlier, the JASON report highlighted under-development of experimental work. The development of the usable privacy and security community (SOUPS) signals progress here—e.g., usable security papers often show that actual user behavior deviates enormously from what was modelled or assumed. For example, the work by Whitten and Tygar [53] and by Schechter et al. [54] were influential in demonstrating that email encryption tools and browser security indicators were not as easily understood as their designers believed. Articles by Shostack [55] and Maxion [56] in TNW (see Appendix D) emphasize data gathering and the importance of good experimental method. Peisert and Bishop [57] stress the importance of clearly-stated hypotheses and good experimental design; they are more optimistic than others on the applicability of the scientific method to security.

*3) **Metrics and measurement**:* Efforts to define security metrics [58] [59] support the empirical camp. Progress here has been slow; Pfleeger [60] suggests that we are learning from our mistakes, an important step forward being to "stop insisting that quantitative is better than qualitative; both types of measurement are useful" (see also [61], [62]). Sanders suggests value in relative metrics [63]. Stolfo et al. [64] describe many challenges and offer ideas on how metrics might advance our field. A survey by Verendel [65] finds that despite significant work, little evidence supports the hypothesis "security can correctly be represented with quantitative information", and notes:

> *Quantified security is thus a weak hypothesis because a lack of validation and comparison between such methods against empirical data. Furthermore, many assumptions in formal treatments are not empirically well-supported in operational security and have been adopted from other fields.*

While a major challenge is to measure the right things, i.e., those that will convey important aspects of security, "things that are not observed such as new attack approaches are not going to contribute to metrics. It is not possible to definitively measure a level of security" [48, p.4].

*4) **Scientific training**:* Longstaff et al. [66] (see also [67], Appendix D) argue that many computer security researchers, e.g., whose background is Computer Science or Mathematics, have been exposed to little training in experimental science or scientific methods; and that what is needed is better knowledge of scientific methods, both for researchers planning their own research, and when serving as peer reviewers. Remedies suggested include revisiting curricula, and nurturing a subcommunity which both demands and rewards scientifically executed research in experimental computer security. Appendix D mentions LASER and DETER.

*5) **Attack papers**:* Less formally documented is a view emerging in the community that attack papers are too many, and given too much emphasis. For example, on this issue in a 2016 panel [68], stated views included that perceived urgency to publish attack papers is unfounded [69], and that in an analogy to medicine, publication of attacks does little to "help make the patient better" [70]. Schneider writes [71]: "Although somebody does need to be uncovering these vulnerabilities, we should be careful not to portray the activity as *research*." Negative views on attack papers may arise from attack talks at hacker venues (e.g., Blackhat, Defcon), as different goals are perceived between academic research and the "cons." Independent of attack papers at academic venues, Bratus and others [72], [73] argue for learning from hacker culture, publications, and offensive security (cf. USENIX WOOT).

Strong differences in opinion often arise on the value of raising awareness about vulnerabilities. It is thus important to distinguish between *vulnerability papers* demonstrating specific exploits enabled by implementation errors, and papers showing entirely new classes of attacks, triggering a revisiting of beliefs and understandings, and a rethinking of architectures and defenses. For example, it was commonly assumed that preventing injection of malicious code prevented malicious execution—but the generalization of return-to-libc attacks to return-oriented programming [74], [75] showed that malicious execution can result from stringing together segments of legitimate code combined with altered control flow. Basin and Capkun [76] argue for the research value of attack papers, e.g., when they provide broad new insights.

D. Is Security Special?

Feynman makes doing Science sound easy (Section II-D): guess laws, compare predictions to experiment and what disagrees with experiment is wrong. This simple recipe gets complex on considering the details. If we lack even guesses at fundamental laws in security, how are we to proceed?

It is sometimes argued that security has special difficulties and unique challenges that preclude placing the field on a more scientific footing. For example: "The challenge in defining a science of cyber-security derives from the peculiar aspects of the field" [48]. As noted in Section III-B, others question whether a Science of Security is even possible [46]. We review some objections and then respond to them.

*1) **Adaptive adversary**:* That security faces an adaptive, intelligent adversary is often singled out. Geer writes [44]: "But let me be clear about one thing that may make cybersecurity different than all else and that is that we have sentient opponents." While bridge-builders must address hostile conditions, Nature does not form new types of storms in response to improved techniques. Security is seen as uniquely difficult.

*2) **Absence of invariant laws**:* Some observe that many findings, e.g., in Physics, are expressible as invariant laws. Another view is that it is naive to expect a security equivalent to Newton's laws or Maxwell's equations. For example [44]: "There is something different about a search for truth when there isn't any, or at least any that lasts long enough to exhaustively explore." JASON shares this view [48, p.16]:

> *It is unlikely that the science of cyber-security will look much like the universal truths of general relativity and we conclude that because of its overall artificial construction, there are no 'fundamental' laws of cyber-security as ... for example in physics.*

Similarly, Evans and Stolfo write [46]: "Computer security is too entwined with human behavior and engineered systems to have universal laws at the physics level."

*3) **Man-made artifacts**:* Quite apart from adaptivity of the attacker, Computer Security must deal with constantly and rapidly evolving conditions, including evolving hardware and software technology. JASON notes [48] "cyber-security is an artificially constructed environment that is only weakly tied to the physical universe" and "the threats associated with cybersecurity are dynamic." As one example, conclusions for one vendor's browser often do not hold for others, as Zalewski demonstrates in detail [77]. Moreover, results for a given vendor's browser today may differ from results for last year's (or last month's) version—such is the challenge when dealing with a fast-changing science of the artificial (cf. Section II-F).

*4) **Against the view security is special**:* We point out that definitions of Science are intentionally independent of details of the discipline under study. Popper and later philosophers sought a demarcation criterion of use whether investigating General Relativity, heredity in fruit flies, Marxist theory or phrenology—without specific pre-conditions, e.g., that a discipline have invariant laws or be free of active adversaries. Indeed the logical positivists argued that a scientific approach, such as the hypothetico-deductive model, was simply the most reliable way of investigating matters of fact—we might be unhappy with the constraints it imposes or the strength of statements it allows, but no clearly superior alternative is available. Arguing that science is inappropriate for Security appears to require arguing that we have no need of its essential elements: consistency and self-correction (Section II-E).

Biological and military systems must also guarantee robustness in the presence of adversaries (see Forrest et al. [78], [79]). In that many of its discoveries are expressible as laws, Physics is an exception rather than the rule [3, pp.197-208]. The compactness and elegance of the mathematical expression of much of Physics is unmatched in other Sciences—e.g., "most biology has little use for the concept of a law of Nature, but that does not make it less scientific" [4, pp.200-201].

Numerous branches of Science have overcome difficulties that once seemed unique and insuperable. Pleading uniqueness to avoid being held to scientific approaches is common in unscientific fields, and would place Security in poor company.

E. Crypto Imperfect: Definitions, Models, Proofs, Real World

Cryptography is held up as a role model for Science in Security, e.g., as implied by Schneider's view in offering a blueprint [51] (quoted in Section III-C). Without doubt, cryptographic research has had many successes. Yet no area is perfect, and specifically because of the high status it enjoys, here we focus on internal challenges within cryptographic research itself. While it is not our intention to take sides or pass judgement, in what follows, examples are chosen to specifically highlight community divisions, and topics on which there is lack of consensus. Some researchers may be sensitive to issues raised; our view is that it is precisely such issues that offer the best learning opportunities, and we encourage further discussion through peer-reviewed literature. This is especially important when selecting role models for a Science of Security—an issue we return to in Section V.

1) ***Reductionist proofs and crypto-as-science challenged****:* "Provable security" involves proofs showing that breaking a target cryptosystem allows solving a believed-hard problem in not much further effort (i.e., "not much" in the asymptotic sense). Bellare [80] suggests *reductionist security* as a less misleading term. More strongly, Koblitz and Menezes [81] note that "proof" and "theorem" historically imply 100% certainty while provable security results are highly conditional, and suggest researchers "should strip away unnecessary formalism, jargon, and mathematical terminology from their arguments"; they go on to criticize crypto researchers who recommend specific real-world parameter sizes despite known limitations of complexity-theoretic proofs—e.g., in extreme examples [82], recommended parameters yield meaningless proofs such as bounding attacker time to be a non-negative number, or greater than a tiny fraction of one second. They also note that since reductionist proofs are strongly tied to specific attack models, nothing can be said about attacks outside of the models/assumptions underlying the proofs. Section V returns to the issue of limitations of models.

Against the view of crypto as role model, Koblitz and Menezes declare crypto itself as far from Science [81]: "The history of the search for 'provable' security is full of zigzags, misunderstandings, disagreements, reinterpretations, and subjective judgements"; in counterpoint, we note that these are not unique in the history of science. Hibner Koblitz et al. [83] note, but contest, prominent cryptographers who "are categorical in their rejection of any notion that cryptography is not fully a science"; and express skepticism observing the use of "high-status terms such as 'science' and 'mathematical proof' that becomes more fervent even as the field is showing itself time and again to be as much an art as a science". In contrast, Degrabriele et al. [84] (see III-E2 immediately below) assert that provable security has transitioned crypto to a science (a view also stated by Katz and Lindell [85, preface]), while acknowledging the gap (cf. [86]) between what proofs promise and deliver in the real world, e.g., due to limited models.

2) ***Provable security vs. the real world****:* Provable security methods, despite their imperfections, can rule out important classes of attacks. Here we discuss other challenges they face in the form of side-channel attacks. These are well-known to be powerful—e.g., Brumley and Boneh [87] demonstrated that private keys of an OpenSSL-based web server can be recovered over a local area network, using Kocher's known timing attack. Real world attackers are of course not physically stopped by mathematical proofs—the proofs' models and assumptions logically constrain theorem pre-conditions, not attackers. Degabriele et al. [84] give a higly accessible exposition of "what can go wrong when systems that have been proven secure in theory are implemented and deployed in real environments". An overview notes they provide "insights on the disconnect between science and engineering" [46]. On gaps between crypto theory and practice, they write [84, p.33]:

> *One of the main issues at stake here is the degree of assurance that provable security provides. Researchers have discovered many attacks on cryptographic schemes that were previously proven to be secure. However, we must consider that no science provides an absolute guarantee of the truth of its results, and that sciences evolve over time.*

Describing an SSH side-channel attack [88] by two of them, they note that the obvious question is [84]: "how we would be able to attack a variant of SSH that was already proven secure." The answer: the security model failed to consider differences in failure modes by which real implementations report errors—here, allowing attackers to send small units of ciphertext to repeatedly extract small amounts of information by observed differences in how the receiving system responds, e.g., dependent on message formatting. They note [84]:

> *It might seem strange that there can be such an obvious discrepancy between the theory and practice of cryptography ... Practitioners might think provable security results provide an absolute statement of security, especially if they're presented in such a manner. When they later discover that a scheme is insecure because of an attack outside the security model, this might damage their confidence in the whole enterprise of provable security.*

A less tolerant view notes this as "strikingly similar" [89, p.20] to the what allowed a 2011 attack on TLS by Paterson et al. [90]; and that it "was due not to a new type of attack, but a clever variant on the same type of attack the security proof was supposed to guarantee against" [89].

Degabriele et al. [84] explain other practical side-channel attacks despite provable security proofs on MAC-then-encrypt constructions, including an SSL/TLS mechanism exploiting observable timing differences caused by padding errors, and an IPsec construction in which formatting differences cause either further processing or a dropped packet. Side-channels continue to epitomize the difficulty of modeling real world problems—raising problems even with definitions. Appendix E discusses this and other crypto-related issues further.

IV. Failures to apply lessons from Science

We now detail security research failures to adopt accepted lessons from the history and philosophy of science.

A. Failure to observe inductive-deductive split

Despite broad consensus in the scientific community, in Security there is repeated failure to respect the separation of inductive and deductive statements. Both types have value, provided their limitations are recognized and they are kept separate. However, we find numerous examples where the two are confused, or limitations of formal statements are ignored.

Schneider [51] suggests modifying the definition of Science to include deductive statements. "The status of the natural sciences remains unaffected by changing the definition of a science in this way. But computer science now joins." We note this is not a minor change. Using falsification as demarcation criterion was not an arbitrary choice. This suggested definition erodes the distinction highlighted in Section II-A, obscuring the point that purely deductive statements cannot describe real-world events.

Speaking of mathematical guarantees as if they are properties of real-world systems is a common error. Shoup suggests that with provable security (quoted in [89]) "we essentially rule out all possible shortcuts, *even ones we have not yet even imagined*. The only way to attack the cryptosystem is a full-frontal attack on the underlying hard problem. Period." This dismisses the fact that it is a real-world system that must resist attack, not a mathematical one. Recalling the comments in Section III-C1 on the superiority of formal approaches (e.g., the quotation from [52]): while it is correct to note that empirical evidence can never demonstrate the infeasibility of attacks, we must also note that the same is true of formal reasoning. The conclusion that formal approaches enjoy an inherent superiority in this respect is unsound. A proof can deliver guarantees only about a mathematical system, not a real-world one. Since it is real-world systems that we ultimately use, the choice is not between one approach which offers immunity to attack and another which does not. Rather, the question is to what degree properties proven about a mathematical system can be translated into useful properties of a real-world one. Speaking of mathematical and real-world systems in the same argument and suggesting that the properties proved of one are naturally enjoyed by the other is a recurring error. JASON [48] acknowledges a research/real-world gap, but fails to highlight this fundamental limitation of formal approaches.

Indeed the term "provable security" (Section III-E) involves overloading or ambiguous use of one of its words. If security is proved in the mathematical sense, then it can't refer to a real-world property (such as the avoidance of harm). Conversely if security refers to a real-world property, it cannot be proved (any proof lies in the deductive realm). Using the term "provable security" (likewise, "proof of security") regularly and reliably generates unnecessary confusion.

Thus "provable security" is not a term that can be taken literally. It might be argued that the limitations are well-understood and the overloading of the term causes little harm. First, this appears optimistic and McLean already warns in 1987 (quoted in Section III-A) of the harm of implicit assumptions "which everyone knows about." Second, we point out that it is precisely when they are intended literally that scientific statements are most useful. Neptune was discovered in 1846 because of a small wobble in the orbit of Uranus. Gravitational waves and the Higgs boson were discovered because of minute deviations from what would have been expected in their absence. While we caution against taking Physics-envy too far, these findings would not have emerged if the underlying theories hadn't been capable of being taken literally. With no wiggle room or implicit assumptions to fall back on, anything not explained by measurement error is a discovery. Thus anything that gets in the way of taking claims literally impedes progress.

Finally, aren't some conclusions so simple that they can be deduced? A simple example may flush out misunderstandings. For example, we might think that the claim "a 128-bit key is more secure than a 64-bit key" can be evaluated purely by deduction. First, we must separate the question of whether we believe a particular claim from the question of whether the reasoning behind it is sound; a valid conclusion doesn't imply a solid argument. If the reasoning here is that "more secure" means "having a longer key" then the claim is simply a tautology. By this definition one key is more secure than another even if both are so short that guessing is trivial, both are so long that guessing is impossible, or an attacker is impervious to length, such as in a key-capture attack. So, "more secure than" doesn't, on it's own, say anything about a possible difference in outcomes. Only when we add a real-world assumption (e.g., presence of an attack for which 128-bits is much more difficult than 64) does it allow real-world conclusions. Similarly, using pure deduction to defend the claim "a system protected by a password is more secure than one without" we will be reduced to making the circular observation that a password blocks those attackers for whom a password is a barrier. Tautologies can be used to justify everything from strong passwords to Faraday cages [91]. If we allow tautologies to dictate resource allocation we must tackle the question of which tautologies to allow and which to reject.

B. Reliance on unfalsifiable claims

There is also considerable failure to avoid unfalsifiable claims and statements. Herley notes [91] that there is an inherent asymmetry in computer security that makes large classes of claims unfalsifiable. We can observe that something is insecure (by observing a failure) but no observation allows us to determine empirically that something is secure (this observation is often used to motivate formal approaches, see, e.g., the remarks quoted in Section III-C1). It follows that claims of insecurity, and of necessary conditions for security, are unfalsifiable. For example, to falsify "in order to be secure you must do X" we would have to observe something secure that doesn't do X. If we interpret "secure" as a real-world property, such as the avoidance of future harm, then observing it requires knowing the future. On the other hand, if "secure" is interpreted formally, while we can now identify mathematically secure systems, we can make no deductions about real-world events (e.g., that harm will be avoided). A similar argument shows that claims of the form "X improves security" are unfalsifiable.

A concrete example may clarify. To falsify the claim "a password must have at least 8 characters and contain letters, digits and special characters to be secure" we would have to find a secure non-compliant password. However, we can't find a secure password, since successfully avoiding harm in the past is no guarantee about the future. The alternative is to formally define security of a password as having a certain structure, or resisting a certain number of guesses, etc. We can of course find necessary conditions if security is defined formally, but these are just restatements of the definition (e.g., a password that withstands 10^{14} guesses is secure if security means withstanding that number of guesses). To relate the formal (e.g., password has certain structure) and real-world (password will not be guessed) notions of security we must make assumptions about what an attacker can and cannot do (e.g., attacker can get access to the hashed password file but cannot execute more than 10^{14} guesses). By symmetry, just as assumptions that attackers cannot do something can never be verified (e.g., there's no way to verify an attacker cannot calculate logarithms in a finite field), assumptions that they can do something can never be falsified.

In summary, claims of necessary conditions for real-world security are unfalsifiable. Claims of necessary conditions for formally-defined security are tautological restatements of the assumptions [91]. Unfortunately statements of this form are commonly used to justify defensive measures, especially where data is rare. Many current practices and beliefs are not based on measurement of outcomes. For example, we have no A/B trials or observations demonstrating improved outcomes for those complying with various password practices, refusing to click through certificate warnings, or "being alert for suspicious links". Rather, these recommendations appear to stem either from authoritarian or unfalsifiable statements. This causes "an asymmetry in self-correction: while the claim that countermeasures are sufficient is always subject to correction, the claim that they are necessary is not" [91]. This can result in accumulation of countermeasures as there is no mechanism for rejecting measures or declaring them no longer required. Thus a failure to avoid unfalsifiable claims may account for the security overload users complain of experiencing [92]–[94].

C. Failure to bring theory into contact with observation

A scientific model is judged on the accuracy of its predictions (Section II-C1); lack of data or difficulty in making measurements does not justify trusting a model on the sole basis of its assumptions appearing reasonable. But this is often done in security research.

Consider for example the long-accepted wisdom that passwords are made stronger by the inclusion of upper-case letters, digits and special characters, recommended by Morris and Thompson [95] to address the observed problem of users choosing English words as passwords, and now widely mandated in part due to a long-standing NIST standard [96]. It was assumed that including digits and special characters would push users to choose random-like strings. Originally, this may have appeared a reasonable assumption (even if false); the strength of users' preference for simple passwords and ingenuity in circumventing security measures was not obvious in 1978. However, storing passwords as salted hashes (a second major recommendation [95]) precluded easily measuring whether mandates on character composition were having the predicted effect. Recent empirical work shows that they do not [97]–[99]. For three decades after 1978, not only were there few apparent attempts to check the accuracy of the prediction, but great effort was devoted to having users follow misguided means to improve password security. Community actions were based on the assumed truth of something that depended critically on an untested assumption.

The non-evidence-based security measures noted in Section IV-B are further examples. The NIST guidelines for authentication [96] acknowledge that, lacking measurements, many of the measures they suggest were justified this way. Apparently, the difficulty of acquiring empirical data in security—e.g., due to instrumentation, privacy, and commercial forces—extends this problem far beyond the examples mentioned. To counter this, Shostack [55] suggests that one step to a Science of Security is a greater emphasis in justifying security mechanisms by pointing to superior outcomes resulting from their use, rather than this being the exception to the rule.

D. Failure to make claims and assumptions explicit

As noted in Section II-B, the evidence falsifying a precise claim is easily described. If a theory says "X should never happen under assumptions A, B and C" then showing that it does suffices to refute the claim. But when a statement is vague, or assumptions implicit, it is unclear what, if anything, is ruled out. Thus, difficulty articulating what evidence would falsify a claim suggests implicit assumptions or an imprecise theory [3].

Consider the large body of work devoted to modifying security-related user behavior (examples noted earlier are paying more attention to TLS warning messages, and choosing passwords). Many large sites, and governments, devote considerable energy to user education. The bulk of this takes the desirability of the goal as given—e.g., that raising awareness of cyber threats or paying more attention to warnings is inherently beneficial. The assumption that this will improve actual outcomes is left implicit and seldom questioned. Examples in the research literature include defining effectiveness as the fraction of users terminating TLS connections after a warning, complying with unvalidated advice on detecting phishing attacks, or choosing a password of a certain format. Many efforts to influence users implicitly assume a goal of minimizing risk. But this implies no measure should ever be neglected; a more realistic goal is to minimize the sum of risk *plus* the associated defensive cost [100]. Unstated assumptions too easily escape debate.

The problem of implicit assumptions seems widespread. This leads Bishop and Armstrong [101] to suggest that the skill of reverse-engineering to uncover assumptions implicit in a security design is a vital part of a computer security education. The problems with attempts to provably deal with side-channels (examined in Appendix E1) offer further examples.

E. Failure to seek refutation rather than confirmation

The limitations of formal approaches noted in Section IV-A might lead to belief that empiricism wins—that measurement and experimentation are the clear way forward for pursuing security scientifically. The truth appears more complex. Recall that in the hypothetico-deductive model (Section II-E), hypotheses are most useful when they allow anticipation of as-yet unseen things, and observations are most useful when they present severe tests to existing hypotheses (vs. simply corroborating existing beliefs). If that model is not to be a random walk, observations must actively seek to refute existing belief (see Section II-D).

Good measurements must severely test existing hypotheses and/or advance the forming of new hypotheses. For example, Rosalind Franklin's X-rays of DNA suggested a twin spiral structure; Galileo's (possibly apocryphal) dropping cannon balls from the Tower of Pisa severely tested existing belief. Recent security-related examples of severely testing existing beliefs are: Zhang et al.'s experiment [102] showing many new passwords easily guessable from old ones after mandated password changes (and thus that assumptions underpinning widespread password expiration practices are false); Weir et al. [103] showing that the commonly used crude-entropy measure for passwords correlates poorly with resistance to guessing; and Bonneau's [97] experiments with and analysis of anonymized password data corresponding to 70 million Yahoo! users, including proposal and use of new partial guessing metrics shown to have greater utility than entropy-based metrics (which can't be estimated even with large samples).

A problem with much empirical work in Security is that it neither generalizes readily (cf. Section II-F) to suggest a new hypothesis, nor presents a *severe test* (see Appendix A) to an existing one. For example, a measurement of a botnet or other phenomenon allows no conclusion about other botnets, without some understanding of how representative it is. A significant difficulty in user studies is guaranteeing ecological validity (without which any observations may not generalize). Do users completing tasks for compensation on a crowd-sourced platform such as Mechanical Turk accurately represent the behavior of typical users? A single paper testing the validity of this assumption (Fahl et al. [98], finding mixed results) is greatly out-numbered by those accepting its truth, and whose conclusions become suspect if it is false.

Much usable security work uses Null Hypothesis Significance Testing (NHST). Here it is the null rather than the alternative hypothesis that we attempt to refute. Thus we end up with evidence in favor of the alternative hypothesis, but never attempt to refute it directly. A reason refutation is emphasized is that confirming evidence generally tells us less than severe attempts to falsify. For example, there may be many hypotheses consistent with rejection of the null hypothesis, not just the particular alternative the experimenter has advanced [104]. The replication crisis in Psychology (see Appendix B) appears at least partially due to a weakness in NHST (in addition to the problems of p-hacking and publication bias).

Section III-C5 noted that many feel there is excessive emphasis on attack research. Partial support for this view is that an attack is a demonstration, or observation, that something is possible, but it is too often unstated what hypothesis this severely tests. As a statement about attacker capabilities, a demonstration that a system is insecure is of interest if it was believed secure; but less so if it was already believed insecure. Alternatively, an attack can be viewed as a statement on defender requirements. A new attack can prove valuable if it provides new insights on what defenders must do, or as noted earlier, corrects false assumptions; papers which demonstrate simple vulnerabilities fail to do that. Demonstrating that a particular side-channel can be exploited gives no signal whether everyone should always adopt countermeasures, some people should, or some should under some circumstances. If it fails to address such questions explicitly, an attack simply insinuates defensive obligations, and viewed this way, offers weak contributions. Scientific papers in other fields expect more than observations alone, e.g., theories or generalizations resulting from observations.

V. Ways forward: insights and discussion

From our review of old science literature and recent security research, can we learn new (or old) lessons, draw insights, find take-away messages? We reflect and offer summary observations and constructive suggestions based on earlier sections.

T1: *Pushes for "more science" in security, that rule nothing in or out, are too ambiguous to be effective. Many insights and methods from philosophy of science remain largely unexplored in security research.*

Reviewing the security literature (cf. Section III-B and Appendix D) for guiding descriptions of Science of Security that don't simply re-use the term, we find few; common are circular exhortations to do things more scientifically. Confusion as to what is (or not) wanted allows every researcher to naturally view their current work as precisely what is needed. Recalling Popper's view that to count as scientific a statement has to "stick its neck out" and be exposed to risk, we suggest that the same is true of pursuing security scientifically: to be effective, calls for more science should specify desired attributes, specific sources of dis-satisfaction with current research, and preferred types of research. Acknowledging current ambiguity may motivate discussion of scientific attributes best serving the community.

T2: *Ignoring the sharp distinction between inductive and deductive statements is a consistent source of confusion in security.*

The importance of this divide, and of being clear which type of statement is being made, is recognized in most branches of Science (Sections II-A, IV-A). It is disappointing then that authors across four decades find it necessary to remind us that it applies to security also [35], [40], [41], [47], [84], [105].

It is worth being unequivocal on this point. There is no possibility whatsoever of proving rigorously that a real-world system is "secure" in the commonly interpreted sense of: invulnerable to (all) attacks. This is not simply because of the possibility of flawed implementation. Formal methods can deduce certain guarantees only if the assumptions are met. However, whether a real-world system meets any set of assumptions is an empirical claim, not something that can be established formally. The combination of a rigorous deductive statement and less-than-rigorous empirical one can never yield a rigorous guarantee. A claim that a real-world system enjoys the same security guarantees as mathematically proven is logically unsound. The value of a formal guarantee is that it concentrates doubt on the assumptions. It greatly simplifies analysis of a real-world system if doubt about its security hinges only on correctness of implementation and attacker ability to solve certain hard problems, rather than on a host of other issues as well. As an example, the fact that Diffie-Helman key exchange relies on a small set of well-understood assumptions allowed Adrian et al. [106] to reverse-engineer where the implementation weakness may lie after a reported NSA compromise.

It is also worth considering why such confusion persists when the literature clearly recognizes the gap between the abstract world of deduction, and the real-world from which science derives facts and makes claims. It seems clear that terms like "provable security" facilitate ambiguity about which side of the inductive-deductive divide we are on: "provable" implies rigorous deduction, while "secure" creates an expectation of real-world guaranteed safety from attack.

T3: *Unfalsifiable claims are common in security—and they, along with circular arguments, are used to justify many defensive measures in place of evidence of efficacy.*

It is important that when defensive measures are recommended they be supported by good justifications. Section IV-B shows that statements of the form "X is necessary for security" or "security is improved if you do X" are problematic: if "security" is interpreted as a real-world property they are unfalsifiable, and if it is interpreted formally they are circular. In contrast, claims about improved outcomes are falsifiable. However, reliable measurements of outcomes are the exception rather than the rule in security. Thus, many defensive measures (including many that are sensible) rely on unfalsifiable or circular justifications.

Unfalsifiable justifications bias towards excess defensive effort: there are many ways to argue measures in, but no way to argue one out. This correlates with the often-observed problem of users overwhelmed with security instructions and advice.

T4: *Claims that unique aspects of security exempt it from practices ubiquitous elsewhere in science are unhelpful and divert attention from identifying scientific approaches that advance security research.*

Several excuses were examined in Section III-D. We point out that negative statements lacking alternatives don't aid progress; actionable statements do. Suggesting that a scientific approach is a poor fit for security, is in no way helpful unless we suggest an alternative that is more appropriate. While security certainly faces major challenges (e.g., changing technology landscapes and intelligent adaptive adversaries), so do other fields. In Astronomy, the paths of planets and stars are not easily controlled as independent variables, but observational experiments prove invaluable; in Life Sciences, the evolution of pathogens changes underlying landscapes; Quantum Physics research continues despite the inability to directly observe subatomic particles. The broadly accepted outlines of scientific method, having evolved over much time and great scrutiny, are by consensus view the best way to figure things out.

Thus, there's little support for the view that science is a luxury. The scientific process described in Section II is not fragile or suited only to problems that possess rare qualities. It simply produces a set of claims that are consistent and self-correcting and that allow anticipation of not-yet-observed real-world events. Falsification as a demarcation criterion is simply an acknowledgement of the fallibility of the process. Relaxing falsification as a criterion requires believing that we never make mistakes in our claims and predictions, or that we are indifferent to the consequences when we do.

T5: *Physics-envy is counterproductive; seeking "laws of cybersecurity" similar to physics is likely to be a fruitless search.*

This observation is not new (e.g., see Section III-D; and [51] for views on laws), but warrants comment. The accomplishments of Physics over the last 150 years may be the most successful scientific research program ever conducted. However, most Sciences do not look like Physics (nor Crypto, below), and we should not pre-judge what a Science of Security will look like. Large sub-areas of Security might be better compared to the Life Sciences [79]. Caution should be exercised that a desire for quantification does not disadvantage the applied or systems research (see T6, T7), or impose mandatory quantitative metrics where no such meaningful metrics are known (see Section III-C3) lest this lead to being "precisely wrong". Admitting the possibility of there being no formal laws to find leaves other paths open. One noted below is to better systematize the "messy aspects" of security.

T6: *Crypto-envy is counterproductive; many areas of security, including those involving empirical research, are less amenable to formal treatment or mathematical role models.*

Without the accomplishments of Cryptography many of the technologies we take for granted might not exist, and it has a special hold on the minds of security researchers. As discussed in Section III-E, it is true that Crypto remains subject to challenges found in many other areas of Science (definitions, confusion due to terminology and language usage, constructing models simple enough to use and yet complex enough to reflect the real world); but that is not unexpected. The main point is that despite many pointing to crypto as role-model for a Science of Security, its methods are less suitable for numerous areas, e.g., systems security and others involving empirical research. Simply wishing for systems security to be as neat and tidy as mathematically-based crypto does not make it so. Crypto's rigorous mathematical foundations are in sharp contrast to, for example, "messy" systems security, and areas that must deal with human factors. Crypto also does not typically involve the type of scientific experimentation found in empirical sciences generally, nor systems security in particular. The formality of Cryptography in no way removes the challenge of creating models relevant to the real world. Crypto does not have a monopoly on benefiting from systematic accumulation of knowledge, clarity and rigor.

T7: *Both theory and measurement are needed to make progress across the diverse set of problems in security research.*

There is little support in the history or philosophy of Science for the view that one or the other holds the key to progress. Major advances in the history of Science have come at various points from serendipitous observations, careful programs of measurement, theoretical insights, and even simple thought experiments. Science uses things we have seen to infer theory that allows us to anticipate and predict things we have not yet seen. The process is iterative, with theory and observation in constant contact. Theory that is not brought into contact with observation risks disconnection from reality and being based on untestable or false assumptions. Indiscriminate measurement offers fewer opportunities for discovery than experiments that deliberately set out to refute or refine existing theory. While both are essential, recent history suggests that theory that has not been tested by observation is currently a greater problem in security than measurement that fails to test theory (see Sections IV-B and IV-C).

T8: *More security research of benefit to society may result if researchers give precise context on how their work fits into full solutions—to avoid naive claims of providing key components, while major gaps mean full-stack solutions never emerge.*

That security research should aim to benefit society is generally accepted, especially when publicly funded [107], [50]. Science has many instances of difficult problems involving complex, multi-part solutions—with the responsibility of ensuring delivery of all parts to a complete solution sometimes taken on by a scientist single-handedly spanning the full spectrum from fundamental research to a fully-engineered solution (cf. Pasteur's Quadrant, Section II-G). It has been observed [108] that security has had challenges in translating academic research from lab to real world. This is of greater concern to agencies viewing technology transfer as their primary objective, than whose main objective is supporting basic science. Regardless, more research of societal benefit may result if researchers took responsibility for explaining how contributions add up to full solutions; these rarely appear by chance. To trigger useful community discussion, one might ask who is responsible for the overall roadmap for emergence of full-stack solutions addressing important problems.

T9: *Conflating unsupported assertions, and argument-by-authority, with evidence-supported statements, is an avoidable error especially costly in security.*

If a security policy is based on authoritarian statements both unsupported by evidence and for which obtaining empirical data is difficult, then overturning the policy is difficult—since vague claims are hard to refute (see Section IV-D), and because of the impossibility of establishing that a defense is *not* necessary (see Section IV-B and T3). Such errors are costly since self-correction [14] (absent entirely for unfalsifiable statements) is lost.

It was common to rely on authoritarian statements before the rise of the scientific method. Whereas a Science of Security would be evidence-based, many password policies are not only arguably unsuited for today's environment [94] but also lack supporting evidence (see Sections IV-C, IV-E). Likewise, science-based evidence supporting use of anti-virus software is thin; statements such as it being "necessary for security" are unfalsifiable (it may stop some attacks, but "necessary" is incorrect; other defenses may be equal or better).

Science reserves the term 'law' (cf. T5) for statements that are most general, having survived the severest of tests—but the status of *any* scientific statement rests solely on the evidence supporting it. There is no other source of authority. 'Rules-of-thumb' are not called 'laws' for many reasons, including that they have not been as rigorously tested, nor as precisely stated; similarly for security 'principles'. The only guide to their utility is the evidence supporting them, and for both, we must be careful that they are supported not only by convincing evidence, but that their relevance is continually re-challenged, e.g., as computer ecosystems evolve. Landwehr notes [109, p.2]: "Before the underlying science is developed, engineers often invent rules of thumb and best practices that have proven useful, but may not always work." That they do not always work raises the question of their pros and cons; they risk being applied in inappropriate scenarios, and confused with authoritarian statements. In summary, scientific statements stand or fall on how they agree with evidence. Calling something a principle, best-practice, rule-of-thumb, or truism removes no burden of providing supporting evidence.

T10: *Despite consensus that assumptions need be carefully detailed, undocumented and implicit assumptions are common in security research.*

Failure to make assumptions explicit is discussed in Section IV-D, and failure to check assumptions in IV-C. Connections between abstractions and the real world (Section II-C) are often unchecked or loose in security, as detailed with cryptographic examples in Section III-E; brilliant and deep results may have little impact in the observable world. Greater care in explicitly detailing and challenging pre-conditions would better illuminate the breadth or narrowness of results.

Recommending that assumptions should be carefully documented seems inadequate. The challenge is not in getting agreement on the importance of doing so, but in establishing why we fall so far short of a goal that few presumably disagree with, and how this might be addressed. One possibility is to find a *forcing function* to make assumptions explicit. As one example (towards a different goal), *Nature* demands that abstracts contain a sentence beginning "Here we show that." Platt [26] recommends answering either "what experiment would disprove your hypothesis" or "what hypothesis does your experiment disprove." By convention, many fields expect explicit hypothesis testing. However, forcing authors to present work in a particular way can be viewed as normative, and is often resisted. One three-year effort to force explicit statements of hypotheses at a workshop was "almost completely without effect, because the organizers had to choose between enforcing the language and holding a workshop without papers" [67]. While documenting all known assumptions will not address the problem of incomplete knowledge [110], failing to explicitly record known assumptions is unscientific.

T11: *Science prioritizes efforts at refutation. Empirical work that aims only to verify existing beliefs, but does not suggest new theory or disambiguate possibilities falls short of what science can deliver.*

In science, there is an expectation to seek refuting observations, as discussed in Sections II-D, IV-E. Corroborating evidence is never definitive, whereas refuting evidence is. Our review reveals the central role of deliberate attempts to test assumptions and refute claims (see quotes from major scientists in Section II-D). It is not that supporting evidence has no value, but the greatest trust is reserved for claims that survive concerted efforts at refutation. Large-scale studies of previously un-measured phenomena naturally have great value when they replace speculation with measurement, and when they challenge assumptions (per examples in Section IV-C). The replication crisis in life sciences (Appendix B) is a warning that empirical corroboration of particular hypotheses must be seen as tentative and relatively weak evidence.

VI. Concluding remarks

From the preceding points, some overall observations emerge related directly to Science. A first meta-observation is that the Security community is not learning from history lessons well-known in other sciences. It is often noted (e.g., [62]) that learning accelerates if we learn from mistakes in other disciplines; arguably, security research is learning neither from other disciplines nor its own literature, and questioning security foundations is not new [109, p.3].

A second meta-observation pertains to those seeing the end-goal of security research being to ultimately improve outcomes in the real world. The failure to validate the mapping of models and assumptions onto environments and systems in the real world has resulted in losing the connections needed to meet this end-goal. A rigorous proof of security of a mathematical system allows guarantees about a real-world system only if the coupling between them is equally rigorous. We have seen repeated failure in poor connections between mathematical systems and real-world ones, and consequent failure of the latter to enjoy properties promised by the former. As discussed, the limitations of models, and challenges due to unclear language and definitions, are among problems identified by many, e.g., McLean [35], Good [40], DeMillo et al. [41], Peisert and Bishop [47], Degabriele et al. [84], Koblitz and Menezes [81], and Denning [105].

As also discussed, many instances of Kuhnian-type crises can be identified in security research—e.g., McLean's System Z [39] threatening the foundational Bell-LaPadula model, and advanced covert channels (cf. later side-channels) threatening information flow models generally [111]; the failing perimeter security model [45]; the paradigm shift from RSA-to-ECC [83], and possible future shift to post-quantum cryptography [112]; complexity-theoretic crypto failing to address side-channel leakage [113]; the potential paradigm shift from pre-authentication to post-accountability [44]. Rather than viewing these as signs of an immature field, we conversely see them as positive signals by Kuhn's view that it is the very existence of everyday paradigms (which are then disrupted by crises) that distinguishes scientific fields from others.

That the Security community is experiencing problems historically well-known in other scientific fields is unsurprising—and perhaps even supports claims of being a Science. What is harder to accept is apparent unawareness or inability to better leverage such lessons. We have noted the absence of consensus in many areas of Security, which some might take as signaling an immature field. Natural tension between researchers in distinct sub-areas or using different methodologies, or between theory and applied researchers, is also common in other fields of research. We do not see these as definitive positive or negative signs of Security being a Science.

On a positive note, one point of consensus is that security research is still in early days. Those who pursue a Science of Security should be cognizant of history—including that progress in science is neither steady nor straight-line. Simply wishing for a Science of Security will not make it happen. What is needed is for security researchers to learn and adopt more scientific methodologies. Specific guidance on what those are, and training in recognizing and using them, may help security research become more scientific.

Acknowledgements. We thank the anonymous referees, and all those who provided feedback on preliminary versions of this paper, including especially AbdelRahman M. Abdou, David Basin, Rainer Böhme, Jeremy Clark, Mohammad Torabi Dashti, Jeremy Epstein, Carl Landwehr, Tom Longstaff, Fabian Monrose, Sean Peisert, and our shepherd Úlfar Erlingsson. The second author is Canada Research Chair in Authentication and Computer Security, and acknowledges NSERC for funding the chair and a Discovery Grant.

References

[1] P. A. Karger and R. R. Schell, “Thirty Years Later: Lessons from the Multics Security Evaluation,” in *Proc. ACSAC 2002*, pp. 119–126.

[2] R. R. Schell, “Information Security: The State of Science, Pseudo-science, and Flying Pigs,” in *Proc. ACSAC 2001*. IEEE, pp. 205–216.

[3] A. F. Chalmers, *What is this thing called Science? (4th edition)*. Hackett Publishing, 2013.

[4] P. Godfrey-Smith, *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press, 2009.

[5] I. Kant, *Critique of pure reason*. Cambridge University Press (translated by Paul Guyer; original version 1781), 1998.

[6] A. J. Ayer, *Language, Truth and Logic*. Dover Publications, New York (unaltered reproduction of 2/e, 1946), 2014.

[7] D. Hume, *An enquiry concerning human understanding: A critical edition*. Oxford University Press (ed. T.L. Beauchamp; original version 1748), 2000.

[8] F. Bacon, *Novum organum*. Clarendon Press (1620 originally), 1878.

[9] J. S. Mill, *A System of Logic, Ratiocinative and Inductive*. Longmans, Green, and Company, 1843.

[10] G. Ellis and J. Silk, “Scientific method: Defend the integrity of physics.” *Nature*, vol. 516, no. 7531, p. 321, 2014.

[11] R. P. Feynman, *“Surely You’re Joking, Mr. Feynman!”: Adventures of a Curious Character*. WW Norton & Company, 2010.

[12] P. Medawar, “Is the scientific paper a fraud?” *The Saturday Review*, pp. 42–43, August 1 1964.

[13] M. Friedman, “The methodology of positive economics,” 1953.

[14] K. Popper, *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 1959.

[15] ——, *Unended Quest: An Intellectual Autobiography*. London: Routledge, 1993, originally titled *Autobiography of Karl Popper*.

[16] ——, “Natural selection and the emergence of mind,” *Dialectica*, vol. 32, no. 3-4, pp. 339–355, 1978.

[17] C. E. Shannon, “The bandwagon,” *IRE Transactions on Information Theory*, vol. 2, no. 1, p. 3, 1956.

[18] C. Hoare and F. Hanna, “Programs are predicates [and discussion],” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 312, no. 1522, pp. 475–489, 1984.

[19] D. Pavlovic, “On bugs and elephants: Mining for science of security,” pp. 23–29, in [114].

[20] C. Ellison, “Ceremony Design and Analysis,” report 2007/399 (2007), Cryptology ePrint Archive, https://eprint.iacr.org/2007/399.pdf.

[21] C. Karlof, J. Tygar, and D. Wagner, “Conditioned-safe ceremonies and a user study of an application to web authentication,” in *Proc. NDSS 2009*.

[22] K. Radke, C. Boyd, J. G. Nieto, and M. Brereton, “Ceremony Analysis: Strengths and Weaknesses,” in *Proc. SEC 2011 (IFIP AICT 354)*, pp. 104–115.

[23] C. A. R. Hoare, “How did software get so reliable without proof?” in *International Symp. of Formal Methods Europe*. Springer, 1996, pp. 1–17.

[24] T. S. Kuhn, *The structure of scientific revolutions*. University of Chicago Press (fourth edition 2012; first edition 1962).

[25] R. Rhodes, *Making of the atomic bomb*. Simon and Schuster, 2012.

[26] J. Platt, “Strong inference,” *Science*, vol. 146, no. 3642, pp. 347–353, 1964.

[27] F. J. Ayala, “Darwin and the scientific method,” *Proc. National Academy of Sciences*, vol. 106, no. Supplement 1, pp. 10 033–10 039, 2009.

[28] F. Darwin, *The life and letters of Charles Darwin*, 1887.

[29] P. J. Denning, “The Science in Computer Science,” *Commun. ACM*, vol. 56, no. 5, pp. 35–38, 2013.

[30] H. A. Simon, *The sciences of the artificial*. MIT Press, Cambridge, MA (third edition; originally published 1969), 1996.

[31] D. E. Stokes, *Pasteur’s quadrant: Basic science and technological innovation*. Brookings Institution Press, 1997.

[32] J. H. Saltzer and M. D. Schroeder, “The protection of information in computer systems,” *Proc. IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975.

[33] U.S. Department of Defense, *Trusted Computer System Evaluation Criteria (TCSEC)*, 1983, informally known as the Orange Book.

[34] P. Karger and R. Schell, “Multics Security Evaluation: Vulnerability Analysis, ESD-TR-74-193, Vol. II,” June 1974, HQ Electronic Systems Division: Hanscom AFB, MA. Original technical report at http://csrc.nist.gov/publications/history/karg74.pdf, updated typesetting at https://www.acsac.org/2002/papers/classic-multics-orig.pdf .

[35] J. McLean, “Reasoning about security models,” in *Proc. 1987 IEEE Symp. Security and Privacy*, pp. 123–133.

[36] ——, “A Comment on the ‘Basic Security Theorem’ of Bell and LaPadula,” *Inf. Process. Lett.*, vol. 20, no. 2, pp. 67–70, 1985.

[37] D. E. Bell, “Concerning ‘modeling’ of computer security,” in *Proc. 1988 IEEE Symp. Security and Privacy*, pp. 8–13.

[38] ——, “Looking Back at the Bell-La Padula Model,” in *Proc. ACSAC 2005*, pp. 337–351.

[39] J. McLean, “The specification and modeling of computer security,” *IEEE Computer*, vol. 23, no. 1, pp. 9–16, 1990.

[40] D. I. Good, “The foundations of computer security: We need some,” essay, 29 September 1986.

[41] R. A. DeMillo, R. J. Lipton, and A. J. Perlis, “Social processes and proofs of theorems and programs,” *Commun. ACM*, vol. 22, no. 5, pp. 271–280, 1979, see also responses in C.ACM 22(11):621–630.

[42] M. Schaefer, “Symbol security condition considered harmful,” in *Proc. 1989 IEEE Symp. Security and Privacy*, pp. 20–46.

[43] ——, “We Need to Think About the Foundations of Computer Security,” in *Proc. NSPW 1993*. ACM, pp. 120–125.

[44] D. Geer, “The Science of Security, and the Future (T.S. Kuhn Revisited),” talk at NSF meeting 6-Jan-2015 and RSA 23-Apr-2015, essay at http://geer.tinho.net/geer.nsf.6i15.txt, shorter version “For Good Measure: Paradigm” in *USENIX ;login:* 41(3):80–84, 2016.

[45] B. Blakley, “The emperor’s old armor,” in *Proc. NSPW 1996*. ACM, pp. 2–16.

[46] D. Evans and S. Stolfo, “Guest Editors’ Introduction: The Science of Security,” *IEEE Security & Privacy*, vol. 9, no. 3, pp. 16–17, 2011.

[47] S. Peisert and M. Bishop, “I am a scientist, not a philosopher!” *IEEE Security & Privacy*, vol. 5, no. 4, 2007.

[48] JASON Program Office, “Science of Cyber-security (JASON Report JSR-10-102),” Nov 2010, http://fas.org/irp/agency/dod/jason/cyber.pdf.

[49] NSA, “Building a science of cybersecurity: The next move,” *The Next Wave*, 2015, vol.21, no.1, https://www.nsa.gov/resources/everyone/digital-media-center/publications/the-next-wave/assets/files/TNW-21-1.pdf.

[50] Executive Office of the President National Science and Technology Council, “Trustworthy cyberspace: Strategic plan for the federal cybersecurity research and development program,” Dec 2011, http://www.whitehouse.gov/sites/default/files/microsites/ostp/fed_cybersecurity_rd_strategic_plan_2011.pdf.

[51] F. B. Schneider, “Blueprint for a science of cybersecurity,” pp. 47–57, in [114].

[52] H. Krawczyk, “Letters to the Editor: Koblitz’s Arguments Disingenuous,” in *Notices of the AMS*. AMS, 2007, p. 1455.

[53] A. Whitten and J. D. Tygar, “Why Johnny can’t encrypt: A usability evaluation of PGP 5.0.” in *Proc. Usenix Security 1999*.

[54] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, “The emperor’s new security indicators,” in *Proc. 2007 IEEE Symp. on Security and Privacy*, pp. 51–65.

[55] A. Shostack, “The evolution of information security,” pp. 6–11, in [114].

[56] R. Maxion, “Making experiments dependable,” pp. 13–22, in [114].

[57] S. Peisert and M. Bishop, “How to design computer security experiments,” in *Fifth World Conference on Information Security Education*. Springer, 2007, pp. 141–148.

[58] A. Jaquith, *Security Metrics*. Pearson Education, 2007.

[59] M. Pendleton, R. Garcia-Lebron, J.-H. Cho, and S. Xu, “A survey on systems security metrics,” *ACM Comp. Surveys*, 2016 (to appear).

[60] S. L. Pfleeger, “Security Measurement Steps, Missteps, and Next Steps,” *IEEE Security & Privacy*, vol. 10, no. 4, pp. 5–9, 2012.

[61] S. L. Pfleeger and R. K. Cunningham, “Why Measuring Security is Hard,” *IEEE Security & Privacy*, vol. 8, no. 4, pp. 46–54, 2010.

[62] G. Cybenko and C. E. Landwehr, “Security Analytics and Measurements,” *IEEE Security & Privacy*, vol. 10, no. 3, pp. 5–8, 2012.

[63] W. H. Sanders, “Quantitative security metrics: Unattainable holy grail or a vital breakthrough within our reach?” *IEEE Security & Privacy*, vol. 12, no. 2, pp. 67–69, 2014.

[64] S. Stolfo, S. M. Bellovin, and D. Evans, “Measuring Security: On the Horizon column,” *IEEE Security & Privacy*, vol. 9, no. 3, pp. 60–65, 2011.

[65] V. Verendel, “Quantified security is a weak hypothesis: a critical survey of results and assumptions,” in *Proc. NSPW 2009*. ACM, pp. 37–50.

[66] T. Longstaff, D. Balenson, and M. Matties, "Barriers to science in security," in *Proc. ACSAC 2010*. ACM, pp. 127–129.

[67] R. A. Maxion, T. A. Longstaff, and J. McHugh, "Why is there no science in cyber science? (panel)," in *Proc. NSPW 2010*. ACM, pp. 1–6.

[68] A.-R. Sadeghi (moderator), R. Anderson, D. Balzarotti, R. Broberg, B. Preneel, A. Rajan, and G. Shannon, "Impact of Academic Security Research: Frogs in Wells, Storms in Teacups, or Raw Diamonds? (panel)," 2016, ACM CCS.

[69] B. Preneel, panelist comment in [68].

[70] A. Myers, comment from floor, by Program co-Chair, in [68].

[71] F. B. Schneider, "Breaking-in research." *IEEE Security & Privacy*, vol. 11, no. 2, pp. 3–4, 2013.

[72] S. Bratus, "What Hackers Learn that the Rest of Us Don't: Notes on Hacker Curriculum," *IEEE Security & Privacy*, vol. 5, no. 4, pp. 72–75, July/August 2007.

[73] S. Bratus, I. Arce, M. E. Locasto, and S. Zanero, "Why Offensive Security Needs Engineering Textbooks," *USENIX ;login:* 39(4):6–10, 2014.

[74] H. Shacham, "The geometry of innocent flesh on the bone: return-into-libc without function calls (on the x86)," in *ACM CCS 2007*.

[75] E. Buchanan, R. Roemer, H. Shacham, and S. Savage, "When good instructions go bad: generalizing return-oriented programming to RISC," in *ACM CCS 2008*, pp. 27–38.

[76] D. A. Basin and S. Capkun, "Viewpoints: The research value of publishing attacks," *Commun. ACM*, vol. 55, no. 11, pp. 22–24, 2012.

[77] M. Zalewski, "Browser Security Handbook," 2008, http://code.google.com/p/browsersec/.

[78] S. Forrest, A. Somayaji, and D. H. Ackley, "Building diverse computer systems," in *Sixth Workshop on Hot Topics in Operating Systems*. IEEE, 1997, pp. 67–72.

[79] S. Forrest, S. A. Hofmeyr, and A. Somayaji, "Computer immunology," *Commun. ACM*, vol. 40, no. 10, pp. 88–96, 1997.

[80] M. Bellare, "Practice-oriented provable security," in *Proc. ISW'97*, 1997, pp. 221–231.

[81] N. Koblitz and A. Menezes, "Another Look at 'Provable Security'," *J. Cryptology*, vol. 20, no. 1, pp. 3–37, 2007.

[82] ——, "Another Look at 'Provable Security'. II," in *Proc. INDOCRYPT 2006*, pp. 148–175.

[83] A. H. Koblitz, N. Koblitz, and A. Menezes, "Elliptic curve cryptography: The serpentine course of a paradigm shift," *J. of Number Theory*, vol. 131, no. 5, pp. 781–814, 2011, http://eprint.iacr.org/2008/390.

[84] J. P. Degabriele, K. Paterson, and G. Watson, "Provable security in the real world," *IEEE Security & Privacy*, vol. 3, no. 9, pp. 33–41, 2011.

[85] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*. CRC Press, 2007.

[86] N. Koblitz, "The uneasy relationship between mathematics and cryptography," *Notices of the AMS*, vol. 54, no. 8, pp. 972–979, 2007.

[87] D. Brumley and D. Boneh, "Remote timing attacks are practical," *Computer Networks*, vol. 48, no. 5, pp. 701–716, 2005.

[88] M. Albrecht, K. Paterson, and G. Watson, "Plaintext recovery attacks against SSH," in *2009 IEEE Symp. Security and Privacy*, pp. 16–26.

[89] N. Koblitz and A. Menezes, "Another look at security definitions," *Adv. in Math. of Comm.*, vol. 7, no. 1, pp. 1–38, 2013.

[90] K. Paterson, T. Ristenpart, and T. Shrimpton, "Tag Size Does Matter: Attacks and Proofs for the TLS Record Protocol," in *Proc. Asiacrypt 2011*, pp. 372–389.

[91] C. Herley, "Unfalsifiability of security claims," *Proc. National Academy of Sciences*, vol. 113, no. 23, pp. 6415–6420, 2016.

[92] A. Adams and M. A. Sasse, "Users are not the enemy," *Commun. ACM*, vol. 42, no. 12, pp. 40–46, 1999.

[93] I. Ion, R. Reeder, and S. Consolvo, "... no one can hack my mind: Comparing expert and non-expert security practices," in *Proc. SOUPS 2015*, pp. 327–346.

[94] W. Cheswick, "Rethinking passwords," *ACM Queue*, vol. 10, no. 12, pp. 50–56, 2012.

[95] R. Morris and K. Thompson, "Password security: A case history," *Commun. ACM*, vol. 22, no. 11, pp. 594–597, 1979.

[96] W. E. Burr, D. F. Dodson W. T. Polk, "Electronic Authentication Guideline," in *NIST Special Publication 800-63*, 2006, http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1_0_2.pdf.

[97] J. Bonneau, "The science of guessing: analyzing an anonymized corpus of 70 million passwords," in *Proc. 2012 IEEE Symp. on Security and Privacy*, pp. 538–552.

[98] S. Fahl, M. Harbach, Y. Acar, and M. Smith, "On the ecological validity of a password study," in *Proc. SOUPS 2013*. ACM.

[99] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur, "Measuring password guessability for an entire university," in *ACM CCS 2013*, pp. 173–186.

[100] D. Florêncio, C. Herley, and P. C. van Oorschot, "Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts," in *Proc. 2014 USENIX Security Symp.*, pp. 575–590.

[101] M. Bishop and H. Armstrong, "Uncovering assumptions in information security," in *Fourth World Conference on Information Security Education*. Springer, 2005, pp. 223–231.

[102] Y. Zhang, F. Monrose, and M. K. Reiter, "The security of modern password expiration: An algorithmic framework and empirical analysis," in *Proc. ACM CCS 2010*, pp. 176–186.

[103] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proc. ACM CCS 2010*, pp. 162–175.

[104] A. Gelman, "Confirmationist and falsificationist paradigms of science," 2014, http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/.

[105] D. Denning, "The limits of formal security models," *National Computer Systems Security Award Acceptance Speech*, http://www.cs.georgetown.edu/~denning/infosec/award.html, Oct. 1999.

[106] D. Adrian *et al.*, "Imperfect forward secrecy: How Diffie-Hellman fails in practice," in *Proc. ACM CCS 2015*.

[107] DHS, "A Roadmap for Cybersecurity Research," Nov 2009, https://www.dhs.gov/sites/default/files/publications/CSD-DHS-Cybersecurity-Roadmap.pdf.

[108] D. Maughan, D. Balenson, U. Lindqvist, and Z. Tudor, "Crossing the 'valley of death': Transitioning cybersecurity research into practice," *IEEE Security & Privacy*, vol. 11, no. 2, pp. 14–23, 2013.

[109] C. Landwehr, "Cybersecurity: From engineering to science," pp. 2–5, in [114].

[110] M. T. Dashti and D. A. Basin, "Security testing beyond functional tests," in *Proc. ESSoS 2016 (Engineering Secure Software and Systems)*, ser. Springer LNCS 9639, pp. 1–19.

[111] P. G. Neumann, M. Bishop, S. Peisert, and M. Schaefer, "Reflections on the 30th Anniversary of the IEEE Symposium on Security and Privacy," in *2010 IEEE Symp. Security and Privacy*, pp. 3–13.

[112] N. Koblitz and A. J. Menezes, "A riddle wrapped in an enigma," *IEEE Security & Privacy*, vol. 14, no. 6, pp. 34–42, Nov/Dec 2016.

[113] S. Micali and L. Reyzin, "Physically observable cryptography," in *Proc. TCC 2004*, pp. 278–296.

[114] NSA, "Developing a blueprint for a science of cybersecurity," *The Next Wave*, 2012, vol.19, no.2, https://www.nsa.gov/resources/everyone/digital-media-center/publications/the-next-wave/assets/files/TNW-19-2.pdf.

[115] ——, "Building a national program for cybersecurity science," *The Next Wave*, 2012, vol.19, no.4, https://www.nsa.gov/resources/everyone/digital-media-center/publications/the-next-wave/assets/files/TNW-19-4.pdf.

[116] C. G. Hempel, "Studies in the Logic of Confirmation (I.)," *Mind*, vol. 54, no. 213, pp. 1–26, 1945.

[117] D. G. Mayo, *Error and the growth of experimental knowledge*. University of Chicago Press, 1996.

[118] D. Pavlovic, "Towards a science of trust," in *Proc. 2015 Symp. and Bootcamp on the Science of Security*. ACM, 2015, pp. 3:1–3:9.

[119] J. Ioannidis, "Why most published research findings are false," *PLoS Med 2(8):e124*, pp. 0696–0701.

[120] R. F. Baumeister, E. Bratslavsky, M. Muraven, and D. M. Tice, "Ego depletion: is the active self a limited resource?" *Journal of personality and social psychology*, vol. 74, no. 5, p. 1252, 1998.

[121] M. S. Hagger, N. L. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. Birt, and M. Zwienenberg, "A multi-lab pre-registered replication of the ego-depletion effect," *Perspectives on Psychological Science*, p. 2, 2015.

[122] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, pp. aac4716:1–aac4716–8, 2015.

[123] S. Hossenfelder, "Does the Scientific Method need Revision?" 2014, https://medium.com/starts-with-a-bang/does-the-scientific-method-need-revision-d7514e2598f3#.h76a6l7zo.

[124] N. Koblitz, "Another look at automated theorem-proving," *J. Mathematical Cryptology*, vol. 1, no. 4, pp. 385–403, 2007.

[125] G. Klein, K. Elphinstone, G. Heiser, J. Andronick, D. Cock, P. Derrin, D. Elkaduwe, K. Engelhardt, R. Kolanski, M. Norrish, T. Sewell, H. Tuch, and S. Winwood, "seL4: formal verification of an OS kernel," in *Proc. SOSP 2009*. ACM, pp. 207–220.

[126] G. Klein, J. Andronick, K. Elphinstone, T. C. Murray, T. Sewell, R. Kolanski, and G. Heiser, "Comprehensive formal verification of an OS microkernel," *ACM Trans. Comput. Syst.*, vol. 32, no. 1, p. 2, 2014.

[127] K. Bhargavan, C. Fournet, M. Kohlweiss, A. Pironti, and P. Strub, "Implementing TLS with verified cryptographic security," in *Proc. 2013 IEEE Symp. on Security and Privacy*, pp. 445–459.

[128] C. Hawblitzel, J. Howell, J. R. Lorch, A. Narayan, B. Parno, D. Zhang, and B. Zill, "Ironclad apps: End-to-end security via automated full-system verification," in *Proc. USENIX OSDI 2014.*, pp. 165–181.

[129] C. Hawblitzel, J. Howell, M. Kapritsos, J. R. Lorch, B. Parno, M. L. Roberts, S. T. V. Setty, and B. Zill, "Ironfleet: proving practical distributed systems correct," in *Proc. SOSP 2015*. ACM, pp. 1–17.

[130] R. Meushaw, "Guest Editor's column," two-page preface to [114].

[131] D. Evans, "NSF/IARPA/NSA Workshop on the Science of Security," Nov 2008, report, Berkeley, CA (presentation slides also available online). http://sos.cs.virginia.edu/report.pdf.

[132] T. Benzel, "The Science of Cyber Security Experimentation: The DETER Project," in *Proc. ACSAC 2011*, pp. 137–148.

[133] C. Collberg and T. A. Proebsting, "Repeatability in computer systems research," *Commun. ACM*, vol. 59, no. 3, pp. 62–69, 2016.

[134] A. Datta and J. Mitchell, "Programming language methods for compositional security," pp. 30–39, in [114].

[135] A. Chiesa and E. Tromer, "Proof-carrying data: Secure computation on untrusted platforms," pp. 40–46, in [114].

[136] D. Maughan, B. Newhouse, and T. Vagoun, "Introducing the federal cybersecurity R&D plan," pp. 3–7, in [115].

[137] F. Chong, R. Lee, A. Acquisti, W. Horne, C. Palmer, A. Ghosh, D. Pendarakis, W. Sanders, E. Fleischman, H. Teufel III, G. Tsudik, D. Dasgupta, S. Hofmeyr, and L. Weinberger, "National Cyber Leap Year Summit 2009: Co-chairs' Report," Sep 2009, 58 pages, http://www.nitrd.gov/fileupload/files/National_Cyber_Leap_Year_Summit_2009_CoChairs_Report.pdf.

[138] J. Wing, C. Landwehr, P. Muoio, and D. Maughan, "Toward a Federal Cybersecurity Research Agenda: Three Game-changing Themes," May 2010, http://www.nitrd.gov/fileupload/files/NITRDCybersecurityR&DThemes20100519.ppt.

[139] R. Meushaw, "NSA initiatives in cybersecurity science," pp. 8–13, in [115].

[140] F. King, H. Lucas, and R. Meushaw, "Advancing the science of cybersecurity with a virtual organization," pp. 20–24, in [115].

[141] T. Longstaff, "Barriers to achieving a science of cybersecurity," pp. 14–15, in [115]; talk transcript (15 Mar 2012) at https://www.nsf.gov/attachments/123376/public/transcript.pdf.

[142] C. Landwehr (moderator), A. Acquisti, D. Boneh, J. Guttman, W. Lee, C. Herley, "How can a focus on "science" advance research in cybersecurity? (panel)," 2012, IEEE Symp. on Security and Privacy.

[143] C. E. Landwehr, "On applying strong inference to cybersecurity science," pp. 11–11, in [115].

[144] J. Bau and J. C. Mitchell, "Security modeling and analysis," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 18–25, 2011.

[145] A. Datta, J. Franklin, D. Garg, L. Jia, and D. Kaynar, "On adversary models and computational security," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 26–32, 2011.

[146] P. Gallison, "How does science advance? (Augustinian and Manichaean Science)," in *First Science of Security (SoS) Community Meeting*, Nov 2012, slides at http://cps-vo.org/node/6418.

[147] F. B. Schneider, "Where are we and how did we get here? (Science of Security: Historical Perspective)," in *First Science of Security (SoS) Community Meeting*, Nov 2012, slides at http://cps-vo.org/node/6414.

[148] J. McLean, "The Science of Security: Perspectives and Prospects," in *Symp. and Bootcamp on the Science of Security (HotSoS)*, Apr 2014, slides at http://cps-vo.org/node/12997.

[149] V. Basili, "What the science of security might learn from the evolution of the discipline of empirical software engineering," in *Symp. and Bootcamp on the Science of Security (HotSoS)*, Apr 2014, slides at http://cps-vo.org/node/12995.

[150] J. C. Carver, M. Burcham, S. A. Koçak, A. Bener, M. Felderer, M. Gander, J. King, J. Markkula, M. Oivo, C. Sauerwein, and L. Williams, "Establishing a Baseline for Measuring Advancement in the Science of Security: An Analysis of the 2015 IEEE Security & Privacy Proceedings," in *Proc. 2016 Symposium and Bootcamp on the Science of Security*. ACM, Apr 2016, pp. 38–51.

[151] P. Grubbs, R. McPherson, M. Naveed, T. Ristenpart, and V. Shmatikov, "Breaking web applications built on top of encrypted data," in *Proc. ACM CCS*, 2016, pp. 1353–1364.

Appendix

A. Problems with Falsification

As noted in Section II-B, while the idea of falsification has been extremely influential, Popper's demarcation problem is not fully settled. Some of the remaining difficulties are sketched here, to give a sense of the problems that modern Philosophy of Science tackles—but these difficulties do not seem to trouble major modern scientists who have written about the scientific method, as they largely see falsification as a litmus test (see Section II-D).

1) ***The raven paradox*** *[116]:* While it is agreed that no number of confirming examples can ever establish the truth of a general claim, it seems natural that every confirming instance might increase our confidence. Every black raven we see would appear to add support to the claim "all ravens are black." A problem with this is pointed out by Hempel [116]: logically the claim "all ravens are black" is equivalent to the claim "all non-black things are not ravens." Thus, observing a white shoe or a green apple is as good as a black raven in lending support to the claim, which seems absurd.

The problem is that falsifiability recognizes observations as either being consistent or inconsistent with a claim. Our intuition that some consistent observations provide much more powerful evidence in favor of a claim than others turns out to be hard to formalize. Addressing this problem, Mayo defines a *severe test* as one that a claim would be unlikely to pass unless the claim were true [117].

2) ***The tacking problem****:* Following Popper and the logical positivists, it might seem that claims that are falsifiable and are supported by confirming examples are solidly scientific. A complication is the ability to make composite claims by tacking on unrelated claims. For example, the claim "all ravens are black" is falsifiable and is consistent with many observations. However, the claim "all ravens are black and the Loch Ness monster exists" also meets this standard—it is falsifiable because the first clause is falsifiable, and it is consistent with observations of black ravens. Thus "tacking on" shows that even simple changes to statements make it difficult to classify them as scientific or not.

One might argue that this is a contrived difficulty. A common practice among scientists is to favor the simplest hypothesis consistent with the observations. It seems difficult to imagine the set of facts for which "all ravens are black and the Loch Ness monster exists" is the simplest explanation. Further, Science generally requires that a hypothesis survive severe tests before it is trusted. Using Mayo's definition [117], checking the color of a raven is not a severe test of the composite claim above.

*3) **Duhem-Quine thesis (holism)**:* While conceptually simple, there is the problem of saying precisely what constitutes falsification of a particular theory. A simple example is the statement "there's a $20 bill under a rock on a planet 1 million light years from here." This is falsifiable in principle, but not in practice. A greater difficulty is that many observations depend on a large number of assumptions—e.g., that equipment is working correctly, that it measures what we think it measures, etc. The holism problem is that it is difficult, or even impossible, to isolate falsification of a single statement from assumptions about the observation. Observations themselves often depend on our existing theory of the environment, i.e., are theory-dependent—e.g., that the earth is stationary was considered an observed fact prior to the work of Galileo [3]. Moreover, what we "see" depends on our experience, as detailed by Chalmers [3]. Thus observations themselves are subject to error; and it can be difficult to determine whether an observation is due to the prediction being false or a problem with some other assumption. "A fossil of a rabbit in the pre-Cambrian rock structure" is sometimes offered as an observation that would falsify evolution. However this assumes certain understanding of how fossils are formed, that our methods for determining the age of rocks is accurate, and that our model of complex organisms never preceding simple ones is sound. The ideal case, of describing an observation that offers a clear binary test of a theory uncontaminated by the risk of other sources of error, can be very hard to achieve.

*4) **Weakness of statements about theories surviving tests**:* Falsification offers a clear criterion for ruling theories unscientific. Even among the philosophers of Science who followed, and often disagreed with Popper, there is considerable consensus that if no possible observation conflicts with a theory, then it is unscientific [3], [4]. This says nothing however about declaring a theory scientific, or saying when we can or should rely on an inductive argument. (For an interesting parallel with security, where things can be ruled insecure but not secure, see [118], [91].)

If we have a falsifiable theory it is natural to suppose that corroborating observations (and the absence of falsifying ones) increase our confidence in the reliability of the theory. It remains quite difficult however to quantify this improvement. Popper gave a clear criterion for when we cannot rely on a theory, but refused to say much about when we can; he refused, for example, even to say that we could "tentatively accept" theories that had withstood severe tests. For many later philosophers and practicing scientists (see Section II-D) this position was unsatisfactory. What words we use to say that General Relativity represents our best understanding of the matter seems of small importance so long as we recognize that we currently have no better one and its status is always subject to overthrow by observation.

B. Science is not perfect

The achievements of Science bring unprecedented authority. The label "scientific" is much sought-after; "non-scientific" is a derogatory label, used to diminish the appeal of a claim. But it is worth noting that even areas and theories now well-accepted within Science have had difficulties. Even fields with solid scientific foundations do not progress smoothly in straight-line fashion. We suggest here that it would therefore be surprising to expect straight-line progress in Security.

Many scientific theories well-accepted at one time were later abandoned, and seem not only "obviously wrong" in retrospect, but even amusing given later understanding—e.g., the explanation of combustion by *phlogiston* theory (things burn because they contain phlogiston, which fire releases to the air); the explanation of heat based on a fluid called *caloric* (which flows from hotter to colder bodies); *corpuscle* theory (all matter is composed of tiny particles); and early explanations of magnetism and light waves. Ironically, such "errors" in past expert theories are more evident to those who study the history/philosphy of science, than to many scientists—in part because textbooks are rewritten to focus precious learning attention on what is correct by current beliefs, and scientific papers stop citing literature which peers have come to perceive as incorrect. This propagates the myth that Science progresses in straight lines.

Errors much closer to our own time also exist. Ioannidis [119] points out that since using $p = 0.05$ ensures one "finding" in 20 will be spurious, and given the large bias toward publishing surprising new results, we easily end up with a majority of published statistically significant findings being due to chance. Indeed, a large fraction of reported findings in some fields cannot be replicated. Ego Depletion [120], regarded as a significant finding in modern Psychology, failed to produce any replication in a multi-year study [121]. An effort to reproduce various findings in Psychology found a significant effect was reproduced in only 40% of cases [122].

Even Physics, the envy of many other branches of science, has had problems—e.g., there is considerable disagreement whether multiverse theories can be considered science [10], [123]. So, we should remember: science is not perfect.

C. Automated verification and automated theorem proving

As noted in Section III-A, Good saw formal verification as the way forward in 1987. But already in 1979, DeMillo et al. [41] complained about calling program verifications "proofs"; they emphasized the importance of human processes to find errors in reasoning, writing: "in the end, it is a social process that determines whether mathematicians feel confident about a theorem" and expressed the view that:

> *because no comparable social process can take place among program verifiers, program verification is bound to fail.*

They acknowledge that while of course errors occur in mathematics as elsewhere:

> *The point is that mathematicians' errors are corrected, not by formal symbolic logic, but by other mathematicians.*

This thus warns against other communities who aspire to be more like mathematicians by using automation. Whereas we have commented earlier on the theory-practice gap related to the deductive-inductive split, they note that the errors in deductive logic should not so easily be dismissed, due to [41, p.273]: "The idea that a proof can, at best, only probably express truth." As noted in Section III-A, Koblitz [124] also gives a negative view of using automated theorem-proving for reductionist security arguments in Cryptography. On the specific point of terminology and language, he writes:

> *The problem with such terms as 'automated theorem-proving', 'computer-aided theorem proving', and 'automated proof-checking' is the same as the problem with the term 'provable security' ... they promise a lot more than they deliver.*

Countering these negative comments on formal verification, others are strongly optimistic about formal methods, including the JASON group report discussed earlier. An emerging view is that for security-critical, highly-leveraged programs of manageable size, its high cost is worth the strong guarantees provided. To this end, substantial progress has been made in recent years, including, e.g., machine-checked verification of the seL4 microkernel [125], [126], the miTLS verified reference implementation of TLS [127], and the recent Ironclad and Ironfleet work [128], [129].

D. Government-promoted "Science of Security" initiatives

As noted in Section III-B, here we review major activities of government-led initiatives to promote Science in Security. An objective is to give a sense of the visible outputs.

The NSA, NSF and IARPA responded to an emerging view "that we do not yet have a good understanding of the fundamental Science of Security" and of the need "to consider whether a robust science of security was possible and to describe what it might look like" [130] by sponsoring a 2008 workshop [131], kicking off a broad initiative to advance a "Science of Security". NSA has sponsored a Science of Security best paper prize since 2013 (noted earlier), published three special issues of its *The Next Wave* (TNW) magazine as detailed below, funded "lablets" on this theme at four universities, and hosted themed workshops in 2012, 2014, 2015 and 2016 (originally a Science of Security Community Meeting, now HotSoS). Among other related events, a workshop series called *LASER* (Learning from Authoritative Security Experiment Results) was conceived in 2011, partially NSF-funded, and focused on repeatable experiments as a path to more scientific security work; since 2004, the DETER Cybersecurity Project [132] has supported scientific experimentation in security. A 126-page 2009 U.S. DHS roadmap for security research [107] identified 11 hard problems.

Selected articles from researchers at the 2008 workshop appeared in *TNW*'s March 2012 cybersecurity issue [114]. These showed a diversity of views on methodological approaches and important problems. Landwehr [109] recounts ancient and recent history related to searching for scientific foundations, notes lacking such foundations need not preclude advancing the state-of-the-practice, and recalls Simon's [30] view of Science (teaching about natural things) vs. Engineering (teaching about artificial things, and the need for a *science of design*). Shostack [55] advocates broader data sharing, more hypothesis testing, and faster reactions to innovative attacks; notes the anti-scientific nature of the security field's tendency to suppress open discussion of known flaws; gives informed discussion of where Science, Security and Engineering intersect; notes several "probably false" hypotheses regarding proofs of security; and encourages focus on budget, outcomes, and comparisons. Maxion [56] emphasizes experiments which are repeatable (produce consistent measurements), reproducible by others (cf. [133]), and valid (well-grounded and generalizable), with focus on confounding errors. He cites Feynman: "Experiment is the sole judge of scientific 'truth'." Pavlovic [19] emphasizes the need to understand the relationship between a system and its environment, and reminds that security is complicated by computer system functions being determined by not only code, but human actions (cf. Section II-C2). He observes that the programs-as-predicates view advocated by Hoare [18] is not as applicable to the world of heterogenous inputs as it is to the well-structured world of algorithms. Datta and Mitchell [134] explain challenges, recent research and open problems in *compositional security*, and their pursuit of general secure composition principles (it is difficult to determine the security properties of systems built from components with known security properties). Chiesa and Tromer [135] pursue secure computation on untrusted execution platforms through *proof-carrying data*. Schneider [51] suggests studying systems the properties of which can be established from first principles.

A second 2012 focussed issue of *TNW* [115] on Science of Security overviews in-progress government programs; it promotes multidisciplinary research augmenting CS, EE and Math with ideas from Cognitive Science, Economics and Biology. Maughan et al. [136] overview the coordinated efforts in U.S. cybersecurity R&D strategy since 2008, and the narrowing of five "prospective game-changing categories" of ideas—including hardware-enabled trust, digital provenance, and nature-inspired cyber health—to three target themes: cyber-economic defenses, moving-target defense, and tailored trustworthy spaces (cf. [137], [138]; [50] adds "designed-in security" as a fourth theme). Meushaw [139] outlines NSA's leadership role in coordinating government departments, including a web-based Science of Security Virtual Organization (SoSVO) [140] serving as a centralized information resource to encourage and strengthen information sharing, collaboration and interaction; outlines an array of complementary supporting programs; and notes that the NSA's Information Assurance research group adopted, as a simplistic definition of cybersecurity science, "any work that describes the *limits of what is possible*." Longstaff [141] (see also [66] and related panel [67]) sees barriers to a Science of Cybersecurity including a community culture favoring quick papers vs. time-consuming efforts typical in experimentally-based fields; a lack of proper scientific training; and culture rewarding novelty and innovative technology over scientific accumulation of knowledge.

A third issue of *TNW* [49] on cybersecurity, in 2015, features papers by funded SoS lablets. Other related efforts include Landwehr's 2012 panel [142] asking "How can a focus on 'science' advance research in cybersecurity?"; see also his brief summary [143] commenting on security subfields which do, or should, leverage Platt's idea of strong inference.

The 2011 NSTC Strategic R&D plan discusses scientific method and characteristics, with a major goal to discover *universal laws* [50, pp.10-11]: "Within ten years, our aim is to develop a body of laws that apply to real-world settings". (Sections III-D and V discuss searching for laws.)

An *IEEE Security & Privacy* special issue on *Science of Security* [46] also triggered by the 2008 workshop, features three papers. Bau and Mitchell [144] outline a conceptual framework for security modeling and analysis, with example applications to a public-key network protocol, a hardware security architecture, and web security (cross-site request forgery). Datta et al. [145] consider compositional security (see above) and reasoning about different adversary models and properties. We discuss the third paper [84] in Section III-E.

HotSoS meeting keynotes were given in 2012 by Gallison [146] (see also Schneider [147]) and 2014 by McLean [148] and Basili [149]; three followed in 2015, and four in 2016. Among other papers in HotSoS proceedings since 2014, Carver et al. [150] use Oakland 2015 papers in an effort to establish a baseline for measuring progress in scientific reporting of research. A paper by Pavlovic in the 2015 HotSoS meeting explores the relation between Science and Security [118]. He observes that, just as in Science we can never be sure that we are right only that we are wrong, in Security "We never are definitely secure; we can only be sure when we are insecure." He proposes approaching trust decisions by hypothesis testing.

E. Crypto is not Perfect

As noted in Section III-E, here we give further examples of challenges within the field of Cryptography. The point is not to single out Cryptography for criticism; rather, since crypto is often prominently proposed as the role model on which to base a Science of Security, we find it helpful to be reminded that even the strongest of sub-fields face challenges, and these may offer among the best lessons.

1) ***Side-channel attacks and leakage-resilient solutions****:* Security definitions—fundamental in all security research—are challenged in another paper in the "Another Look" series questioning the rigor and scientific integrity of crypto research (see http://anotherlook.ca/). Beyond illustrating attacks outside the scope of formal models (this is now clear as a recurring failure), a focus is core definitions made long after known (real) attacks, but not considering them. The authors write:

> *In developing a strong foundation for modern cryptography, a central task was to decide on the 'right' definitions for security of various types of protocols ... good definitions are needed to discuss security issues with clarity and precision.* [89]

But they note much disagreement on "what the 'right' definition of security is" for numerous fundamental concepts. Seven years after Kocher's well-known timing attacks, formal models still failed to address side-channels. Micali and Reyzin [113] then wrote of the "crisis" of complexity-theoretic crypto failing to protect against information leakage; finally new formal models promised *leakage resilience*, but on close examination these again showed a disconnect between security expected and that delivered, due to non-modeled attacks. An argument supporting the new models was desired elimination of *ad hoc* security mechanisms practitioners employ to stop side-channel attacks. The authors countered [89] that in practice there is "no avoiding the need for ad hoc countermeasures". Ironically, one may view the ad hoc countermeasures as further augmented by ad hoc assumptions in the new models: an adversary not being permitted to mount side-channel attacks after seeing target ciphertext, a bounded-retrieval model increasing all parameter sizes and then limiting how many bits attackers may capture, and assuming no key leakage during key generation stages (e.g., attackers being inactive)—these being conditions needed for security proofs. Independent of animosity between individuals [86], another recurring question remains: how to close the gap from model to real-world system (cf. [68]).

2) ***Searchable encryption models vs. the real world****:* A final example involves so-called *BoPET* systems based on multi-key searchable encryption (MKSE), e.g., used for encrypted cloud computations. Grubbs et al. [151] show a formal MKSE model of Popa and Zeldovich, complete with proofs, falls to both passive and active attacks; and a commerically popular BoPET called Mylar, supported by proofs under this formal model, fails in practice. Eerily reminescent of McLean vs. Bell-LaPadula (Section III-A), they summarize [151]:

> *We start by showing that the Popa-Zeldovich security definitions for MKSE do not imply the confidentiality of queries even against a passive server...we construct an MKSE scheme that meets the Popa-Zeldovich definitions but trivially leaks any queried keyword. This shows that the proofs of security for MKSE do not imply any meaningful level of security, but does not yet mean that the actual scheme fails.*

They then show that the actual scheme fails also. They advise:

> *BoPETs need to define realistic threat models and then develop formal cryptographic definitions that capture security in those threat models.*

In contrast, the security definitions used, being inappropriate, provided no guarantees that carried over to the actual system—again highlighting a gap between promises and practical outcomes. We repeat because the error itself is repeated time and again (cf. Section VI): what is proven about an axiomatic system should not be expected to constrain a real-world system unless convincing arguments can be made regarding connections between the abstract system and its counter-part in the observable world. This gives another concrete example, from peer-reviewed literature, of the language "proofs of security" misleading even subject-area experts.