OPINION

Reproducibility failures are essential to scientific inquiry

A. David Redish[a,1], Erich Kummerfeld[b], Rebecca Lea Morris[c], and Alan C. Love[d]

Current fears of a "reproducibility crisis" have led researchers, sources of scientific funding, and the public to question both the efficacy and trustworthiness of science (1, 2). Suggested policy changes have been focused on statistical problems, such as p-hacking, and issues of experimental design and execution (3, 4). However, "reproducibility" is a broad concept that includes a number of issues (5) (see also www.pnas.org/improving_reproducibility). Furthermore, reproducibility failures occur even in fields such as mathematics or computer science that do not have statistical problems or issues with experimental design. Most importantly, these proposed policy changes ignore a core feature of the process of scientific inquiry that occurs after reproducibility failures: the integration of conflicting observations and ideas into a coherent theory.

Here we argue, using examples from mathematics and computer science, that current discussions of the reproducibility crisis overlook the essential role that failures of reproducibility play in scientific inquiry. This viewpoint that reproducibility is a key part of inquiry suggests several new perspectives and policies to promote good science. First, science needs to be given the time necessary to reconcile conflicting results. It typically takes decades to probe the parameter space of a discovery to identify and characterize the fundamental variables. Second, reproducibility failures are a critical part of this journey, and attention

Fig. 1. Discussions about a "reproducibility crisis" often ignore what takes place when reproducibility fails: the integration of conflicting observations and ideas into a coherent theory. Image courtesy of Dave Cutler (artist).

[a]Department of Neuroscience, University of Minnesota, Minneapolis, MN 55455; [b]Institute for Health Informatics, University of Minnesota, Minneapolis, MN 55455; [c]Department of Philosophy, Stanford University, Stanford, CA 94305; and [d]Department of Philosophy, Minnesota Center for Philosophy of Science, University of Minnesota, Minneapolis, MN 55455

The authors declare no conflict of interest.

Published under the PNAS license.

Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and have not been endorsed by the National Academy of Sciences.

[1]To whom correspondence should be addressed. Email: redish@umn.edu.

must be paid to the process of reconciling conflicting results. Third, success or failure should not be based on the conclusions of one or a few studies; strategies such as those of theoretical and synthesis articles that integrate diverse perspectives should be encouraged. We argue that the decades-long process of metabolizing reproducibility failures through theoretical integration is what leads to the reliable results across the sciences that have provided us with remarkable life-changing medical and engineering consequences.

Failure to Generalize

A typical scientific discovery is an observation—in this time and place, under these conditions, an outcome was observed. When another researcher observes a different outcome in a replication study or is unable to reproduce a particular analysis, inquiry should seek to discover the hidden variables that underlie these differences. This was a key advantage of the newly created scientific journal in the 1600s. A formal, public report of an experiment could be better compared and contrasted with other experiments, even if the observations were incomplete. Over time, the scientific journal article made it possible to present observations (even differences in outcomes, i.e., reproducibility failures) that could then be integrated into broader accounts that were more generalizable.

Scientific progress depends on integrating the lessons learned from repeated experiments that produce discordant outcomes. Several researchers have argued that science moves forward by identifying where reproducibility fails (6, 7) and that the key to science is that every answer opens up new questions (8). The identification and interrogation of reproduction failures generates more reliable and enduring scientific knowledge by isolating and characterizing the crucial factors that underlie phenomena.

In many of these cases, what have been called "failures to replicate" are actually failures to generalize across what researchers hoped were inconsequential changes in background assumptions or experimental conditions (9). Moreover, many of these attempts are based on incomplete explanations of the relevant mechanisms and overly fast transitions to clinical practice. Science depends on adequate mechanistic explanations—not just that something works, but how and why it works (10). The amazing past success and ongoing advance of science derive from a cycle of observation, theory, replication, failure, and reintegration, which leads, once again, to new questions, new observations, and new failures (6–8, 10).

This has significant consequences for the practical outcomes of scientific research: The translation of fundamental discoveries into technological innovations or clinical applications takes time. Many commentators assume this translation can occur almost immediately, but the process of inquiry demands sufficient time to measure the space of a discovery (11, 12). The unjustified (but growing) pessimistic attitude from diverse stakeholders (1, 2) toward science comes from a widespread misunderstanding about the role of reproducibility in science.

Importantly, the process of exploring the space of a discovery to find and characterize primary variables in response to reproducibility failures is a component of all genuine inquiry and occurs even in mathematics and computer science, which by definition do not have problems with p-hacking, experimental design, or material practices. Both mathematics and computer science are rife with examples of reproducibility failures that have led to important breakthroughs.

To illustrate this idea in more detail, we examine three examples: the Four-Color Theorem, Fourier analysis, and the development of neural networks. All of these examples proceeded through a multidecade process of examination and inquiry, with boom-and-bust cycles of excitement over the potential of initial results, followed by indications that the results did not live up to expectations, but then followed by the identification of new variables that facilitated a deeper understanding of the original topic.

The Four-Color Theorem

Even after a proof is published, mathematicians continue to check that it is correct, i.e., they attempt to "reproduce the proof." Sometimes they fail: A proof that was previously accepted turns out to be incomplete or contain an error. This happened with an early proof of the Four-Color Theorem. The Four-Color Theorem states that for any map drawn in a 2D plane and divided into contiguous regions, four colors suffice to color these regions so that no two adjacent ones are given the same color. It was first conjectured by Guthrie in 1852. Twenty-seven years later, Kempe (13) announced that he had proven it. However, 11 years after that, Heawood (14) uncovered a fatal flaw in Kempe's argument. Heawood's attempts to reproduce Kempe's proof had failed. Yet Heawood did more than point out the error in Kempe's work: He demonstrated that Kempe's methods could be used to establish a weaker Five-Color Theorem. Over the next century, other mathematicians refined and extended the innovative ideas that Kempe's work contained, which eventually contributed to Appel and Haken's 1976 computer proof of the Four-Color Theorem (15).

In many of these cases, what have been called "failures to replicate" are actually failures to generalize across what researchers hoped were inconsequential changes in background assumptions or experimental conditions.

However, the story doesn't end there. Some mathematicians were skeptical of Appel and Haken's computer proof, both because it was complicated and in light of the invalidation of Kempe's original proof. So mathematicians reproduced the theorem by proving it in a different way. Robertson and colleagues (16) obtained a simpler and more efficient computer proof of the theorem in 1996, and Gonthier (17) formalized it and verified its correctness in 2005. When faced with reproducibility failures in the form of an invalid proof

and questions about the validity of particular methods, mathematicians sought to better understand these methods, results, and inferences over time, which led to new mathematical techniques and different ways to prove the Four-Color Theorem.

Fourier Series

Mathematics also experiences "failures to replicate" in the form of overgeneralizations. For example, in 1807, Fourier made a grand claim: *every* function can be represented as a series of sines and cosines (i.e., as a Fourier series). However, Fourier's initial work was rejected by the French National Academy and dismissed by eminent mathematicians of his time, which delayed the publication of his ideas until 1822 (18).

Despite its negative reception, Fourier's work contained important insights that sparked the interest of mathematicians such as Dirichlet, who in 1829 published a proof showing that Fourier's theorem was correct under limited conditions, but who also presented a counterexample disproving Fourier's general theorem (19). This counterexample was the first so-called "pathological" function, and over the next 100 years mathematicians found more examples of strange "functions" with bizarre and surprising properties (20).

The discovery of these functions violated the mathematical intuitions of the time, leading mathematicians to develop more rigorous and powerful analyses to reintegrate these functions into mathematical theory. This culminated in the development of a theory of "generalized functions" or "distributions" by Schwartz and others in the 1950s (21, 22). These functions are extremely useful in a variety of contexts, such as quantum physics (e.g., the delta function) and electrical engineering (e.g., the impulse function) (23). Today, the work initiated by Fourier and Dirichlet is the cornerstone underlying spectral analysis, which is used in dozens of fields, ranging from electrical engineering to physics and neuroscience.

From Neural Networks to Deep Learning

In both the mathematical and computational sciences, progress depends on a long cycle of breakthroughs, recognitions of the limitations of those breakthroughs (i.e., failures to replicate and generalize), and reintegration leading to new discoveries. In 1958, Rosenblatt proved the Perceptron Convergence Theorem, which opened up the possibility of using "subsymbolic" neural network representations in which information was distributed over many small units as computational learning engines (24). The perceptron is a computational device that sums a set of inputs and applies a nonlinear threshold. Rosenblatt's proof demonstrated with mathematical rigor that these computational devices could learn pattern recognition directly from examples.

At the time, this result was taken to imply that perceptrons could enable computers to think and thus serve as a model of animal learning processes. However, 11 years later, Minsky and Papert (25) published a proof that perceptrons could not solve the simple parity operation. Parity is a logical operation that counts the number of 1s and 0s in a list of 1s and 0s and returns 1 if the count is odd and 0 if even. The simplest example of parity is the XOR problem (a XOR b = true if the true/false statements a and b are different). That perceptrons could not solve parity was a devastating blow to the field and treated as a failure inherent to the nature of the subsymbolic, pattern-completion architecture exemplified by perceptrons.

However, in 1986, it was found that a multilayer network of perceptrons could solve XOR (26). Rumelhart and colleagues (26) demonstrated a method for backpropagating errors across neural networks with hidden layers, which enabled multilayer networks to do the sort of pattern-learning exhibited by perceptrons. This led to the development of a variety of neural network algorithms over the next decade, but these results were again found to be of limited use for complex problems: The computational tools available at the time could not implement these more complicated networks and required too much data to train them.

Over the next 20 years, additional insights into learning algorithms, the increase in computational power due to parallel computing, and the explosion of Big Data provided the necessary components to make these networks useful. Today, these multilayer perceptron-based networks (now called "deep learning") are a highly successful computational paradigm that underlies practical solutions to diverse pattern-recognition problems, including the ubiquitous handwriting, facial, and spoken-language recognition that we use on a routine basis (27).

Success Requires Failures

These examples demonstrate that all inquiry is a dynamic exploration of the space of a discovery in response to failures of reproducibility. Iterated comparisons and contrasts of methods, results, and inferences facilitate integration into robust, generalized accounts. In a very real sense, science is a journey, and no article should be seen as the final answer to a question; every article opens up new questions (8), and many of these new discoveries come directly from failures to replicate (6). Although good experimental design and data management are obviously important parts of conducting good science (4), these examples show that failures of reproducibility occur even in fields in which these specific problems do not arise. This implies that reproducibility failures should not undermine the efficacy or trustworthiness of science. In contrast, they tell us something very important about how scientific inquiry works and how science makes progress (6–8, 10).

Recognizing the centrality of failures of reproducibility to inquiry not only shifts our entire perspective on the reproducibility crisis but also has important policy implications, which differ from those typically discussed (1–5).

First, and most importantly, scientific conclusions should never be based on single studies. The recursive interrogation of methods, results, and inferences with respect to properties such as robustness and generalizability involves multiple studies. It is this process that underlies successful translation into

practical outcomes. Strategies that nurture this interrogation, such as theoretical studies and synthesis articles, which integrate results from different perspectives, should be encouraged.

Second, attention needs to be paid to the activity of reconciling conflicting results. Researchers need to recognize that reproducibility failures are a normal part of science and do not necessarily indicate incompetence or fraud. Initiatives and policies that attempt to curb failures of reproducibility miss the need for strategies that metabolize them, such as utilizing multiple methods to confirm a result and to find the fundamental variables underlying a phenomenon.

Third, science needs to be given sufficient time to reconcile conflicting results. Progress from scientific breakthrough to medical or engineering consequences that change lives does not occur overnight. Probing the parameter space of a discovery, identifying how to control those parameters, and actually making something work takes time. The journey from the discovery of the backpropagation of error-learning rules to deep-learning networks embedded in every smartphone took 30 years. This timeline is typical of many fields of research (11, 12).

A longer timeline is critical because failures when exploring a scientific question lead to new discoveries, but failures in application (such as in clinical trials) can lead to dangerously negative consequences. It is one thing to find that a behavioral manipulation differs across strains of rats (28). It is another to find that an anti-nausea drug given to pregnant women produces limb deformities in their children (29). Rushing the process of translation without a solid understanding of the critical variables is inherently dangerous. Often, decades are necessary because science requires time to chart the "warranty space" of results: What are the parameters across which the discovery is valid? What are the mechanisms underlying those factors? It takes time for replications to fail and time to reintegrate those failures into coherent theories.

The widespread dissemination of this perspective to researchers, research funders, and the general public could positively influence the trajectory of modern scientific practice by preventing overzealous negative responses to the perceived reproducibility crisis and instead redirect efforts and resources to generate more reliable and enduring knowledge. For example, funding agencies need to revise policies to address these recommendations. Studies that pursue the interrogation of methods or results after discoveries are made should be prioritized. The identification and reconciliation of conflicting results should be a valued component of grant proposals. Expectations for the time required to pursue these activities need to be calibrated explicitly in the funding process. Additionally, communicating this image of science as a journey in educational contexts and media outlets is a critical step for the public to achieve a better understanding of how scientific research actually operates.

Over the course of decades, science leads to remarkably reliable results. This reliability has given us an understanding of evolution and climate change, instantaneous international communication, efficient heating and cooling in our houses, and medical and engineering solutions that have extended our lifespans. We have robots on Mars and have flown probes past Pluto, while our phones can recognize our faces and translate our texts. Just as the discoveries that Kempe's proof was incorrect, Fourier's claim was an overgeneralization, and perceptrons could not solve parity problems eventually led to new insights and stimulated further developments, failures of reproducibility are the raw material of genuine inquiry.

Many of the current concerns about reproducibility overlook the dynamic, iterative nature of the process of discovery where discordant results are essential to producing more integrated accounts and (eventually) translation.

The discovery that an experiment does not replicate is not a lack of success but an opportunity. Many of the current concerns about reproducibility overlook the dynamic, iterative nature of the process of discovery where discordant results are essential to producing more integrated accounts and (eventually) translation. A failure to reproduce is only the first step in scientific inquiry. In many ways, how science responds to these failures is what determines whether it succeeds.

1 Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454.
2 Fanelli D (2018) Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci USA* 115:2628–2631.
3 Munafò MR, et al. (2017) A manifesto for reproducible science. *Nat Hum Behav* 1:0021.
4 Leek JT, Peng RD (2015) Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proc Natl Acad Sci USA* 112:1645–1646.
5 Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean? *Sci Transl Med* 8:341ps12.
6 Firestein S (2015) *Failure: Why Science Is So Successful* (Oxford Univ Press, London).
7 Kuhn T (1962) *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago).
8 Firestein S (2012) *Ignorance: How It Drives Science* (Oxford Univ Press, London).
9 DrugMonkey (2018) Generalization, not "reproducibility." Available at drugmonkey.scientopia.org/2018/02/26/generalization-not-reproducibility. Accessed April 16, 2018.
10 Ben-Ari M (2005) *Just a Theory* (Prometheus Books, Amherst, NY).
11 Contopoulos-Ioannidis DG, Alexiou GA, Gouvias TC, Ioannidis JPA (2008) Medicine. Life cycle of translational research for medical interventions. *Science* 321:1298–1299.
12 Lithgow GJ, Driscoll M, Phillips P (2017) A long journey to reproducible results. *Nature* 548:387–388.

13 Kempe A (1879) On the geographical problem of the four colours. *Am J Math* 2:193–200.
14 Heawood PJ (1890) Map-color theorem. *Q J Pure Appl Math* 2:332–338.
15 Appel K, Haken W (1989) *Every Planar Map Is Four Colorable* (American Mathematical Society, Providence, RI).
16 Robertson N, Sanders D, Seymour P, Thomas R (1997) The Four-Colour Theorem. *J Comb Theory B* 70:2–44.
17 Gonthier G (2008) Formal proof—The Four-Color Theorem. *Not Am Math Soc* 55:1382–1393.
18 Fourier J (1822) *Théorie Analytique de la Chaleur* (Firmin Didot, Paris).
19 Dirichlet J (1829) Sur la convergence des séries trigonométriques qui servent à représenter une fonction arbitraire entre des limites données. *J Reine Angew Math* 4:157–169.
20 Kleiner I (1989) Evolution of the function concept: A brief survey. *Coll Math J* 20:282–300.
21 Gelfand I, Shilov G (1964) *Generalized Functions* (Academic Press, New York), Vol 1.
22 Schwartz L (1966) *Théorie des Distributions* (Hermann, Paris).
23 Dirac P (1930) *The Principles of Quantum Mechanics* (Oxford Univ Press, London).
24 Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* 65:386–408.
25 Minsky M, Papert S (1969) *Perceptrons: An Introduction to Computational Geometry* (MIT Press, Cambridge, MA).
26 Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1: Foundations, eds Rumelhart DE, McClelland JL (MIT Press, Cambridge, MA), pp 318–362.
27 LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
28 Rex A, Sondern U, Voigt JP, Franck S, Fink H (1996) Strain differences in fear-motivated behavior of rats. *Pharmacol Biochem Behav* 54:107–111.
29 Vargesson N (2015) Thalidomide-induced teratogenesis: History and mechanisms. *Birth Defects Res C Embryo Today* 105:140–156.