THE FUNDAMENTAL PRINCIPLE OF PROBABILITY: RESOLVING THE REPLICATION CRISIS WITH SKIN IN THE GAME

HARRY CRANE

If you wouldn't knowingly invest in a hedge fund whose manager is compensated the same for generating big returns as for going insolvent, or place a bet with a bookie who doesn't pay up when he loses, or believe a journalist who doesn't vet his sources, why would you trust claims made by scientists who have little or nothing to lose when their assumptions are wrong, analyses are flawed, or findings are false?

The businesses of investing, bookmaking, and news reporting are by no means perfect, but at least they have built-in mechanisms to align the interests of producer (fund manager, bookie, journalist) and consumer (investor, gambler, reader). Fund managers, bookies, and journalists have to risk their own wealth, reputation, and credibility so that they have funds to manage, books to keep, and audiences to write for. The same goes for doctors, butchers, chefs, and (non-union) plumbers. But why not scientists?

Rather than being evaluated against objectively established standards (e.g., profit, honor, and truth in the above examples), professional scientists police themselves with codes of conduct, peer review, and editorial boards, affording themselves a great deal of upside—in the form of job security, a comfortable salary, millions of dollars in taxpayer-funded grants, and even the possibility of fame and fortune if one of their discoveries hits the mainstream—with the potential for downside mostly limited to exceptional circumstances, such as fraud or misconduct. This asymmetry combined with academia's 'publish or perish' culture leads to a perverse incentive structure in which quick-and-dirty is rewarded more than careful-and-correct, which in turn has led to the so-called scientific 'replication crisis'.[1]

Ioannidis called attention to the adverse effects of this system more than a decade ago in an essay titled "Why Most Published Research Findings Are False" [Ioa05]. More than a decade later, there are numerous proposals on the table to reform the way scientific inquiries are conducted, evaluated, and reported, and with it improve the reliability of published research. But even with widespread interest in reform, the replication crisis shows little sign of receding, as one recent effort successfully replicated only 37% of results published in the psychology literature [Col15].

These problems will persist until science, like investing, bookmaking, and news reporting, implements a built-in mechanism that aligns the career success of scientists with the reliability of their results. I argue here that this is much easier to achieve

[1] See, e.g., https://thewire.in/science/replication-crisis-science for an overview of the replication crisis.

than it may seem, as it requires little or no change in the way scientific inquiries are currently conducted and only minor modification in how scientific conclusions are reported. The key idea is to make conclusions reported in the scientific literature more *scientific*, in the sense of being testable and falsifiable, rather than being merely transparent or reasonable, as evaluated against a rubric of best practices or a checklist of state of the art methods. So instead of shying away from the perceived pitfalls of what has been deemed a 'results-oriented' approach to scientific reporting, I argue for more emphasis on results, by which I specifically mean *actual* results, e.g., clearcut criteria by which to verify or falsify published claims, instead of *reported* results, e.g., statistical significance. It is not enough to police methods, as (well-intentioned) proposals such as Registered Reports (RRs) seek to do [Cha14]. The determination of which methods are 'sound' will be made by the same editors and referees who currently decide whether a result is 'significant', changing not the end goal (publication) or the mechanism (peer review) but rather the means by which that end is achieved (i.e., by applying 'sound methods' instead of reporting 'significant results'). So while such implementations may change specific behaviors, e.g., by eliminating P-hacking and other forms of results-oriented misconduct, they will do so without any guarantee of eliminating the replication problems caused by those behaviors.

The importance of correct scientific *results* over sound scientific *methods* underscores the larger societal role of science, a role which is downplayed by Registered Reports and similar proposals to shift emphasis from results to methods. Unlike purely academic disciplines such as philosophy, history, and mathematics, whose immediate impacts do not extend far beyond the ivory tower, results published in the scientific literature have direct consequences on the economy, society, medicine, business, industry, and public policy. So while focusing on methods may be a prudent approach to ensure consistent science in the aggregate, the correctness of individual studies matters to consumers who base policy and business decisions on specific conclusions in the scientific literature. The fact that an errant conclusion in cancer research, education policy, or economics was obtained using 'state of the art' methods and approved by 'expert' peer reviewers is no consolation to the cancer patient, special needs child, or impoverished citizen affected by misguided policies based on those flawed reports. For regardless of whether an individual scientist knowingly commits fraud, unknowingly misapplies a statistical method, or unluckily draws an errant conclusion due to random error, the end result is the same: the conclusions reported in the peer-reviewed, scientific literature are less reliable, and whoever depends on the reliability of scientific research (i.e., just about everyone) is worse off as a result.

This is not to suggest that scientists must be 100% certain of a conclusion before they report it, as trial and error is critical to scientific progress, but consumers of those conclusions, which includes fellow scientists as well as members of the general public, should be assured that what they are reading is very likely correct, not just because the scientific community vouches for the soundness of the methods or the reasonableness of the conclusions but because the scientists asserting these claims have incentives for being right and face consequences for being wrong. This same basic tenet keeps fund managers, bookies, journalists, doctors, butchers, chefs, and plumbers (except those protected by the union) on their toes, and would do the same for scientists.

As long as the primary objective of *scientists* (i.e., publication) is at odds with the objective of *science* (i.e., the attainment of new knowledge and insights), the reliability of the scientific literature will be influenced more by sociological than by scientific factors. To align the interests of consumers and producers of science, it is not enough to merely eliminate incentives for certain kinds of misconduct or to incentivize specific best practices. The consistent production of unreliable science (intentional or not) must become an unsustainable strategy for a scientific career, and conversely the consistent production of reliable science must become the clearest path to a successful scientific career. Put another way, the academic world that scientists inhabit needs to make contact with the real world in which their conclusions have influence.

Academia vs. Reality. Improving the reliability of published science goes hand-in-hand with aligning the goals of *academic* science and *real* science. Real scientific conclusions can be tested against reality, while academic science can only be evaluated by other academics. Thus, testability marks the line between academic and real science, and charts the logical path toward improving the reliability of scientific literature. Because the peer review decisions that guide publication in scientific journals are based primarily on academic criteria, it should come as little surprise that some of academic science fares rather poorly when graded against reality; see, e.g., the replication study in [Col15].

A key step in distinguishing the academic from the real lies in a clearer understanding of how scientific conclusions are often assessed, by statistical methods and their accompanying probabilities, and how these probabilities provide an organic mechanism by which to evaluate scientific claims, all without any need for peer reviewed journals, expert referees, or bureaucratic oversight. Indeed, a major contributing factor to the replication crisis is that the scientific literature makes no distinction between academic and real probabilities, a distinction critical to understanding the role played by probability in scientific discourse.

The key point is that *real* probabilities can be tested in a way that *academic* probabilities cannot be.

> *Academic probabilities are theoretical calculations with only a hypothetical connection to the real world.*

Academic probabilities are the result of armchair analyses, as when a Bayesian philosopher explains probability in terms of a hypothetical disposition toward betting, or when a scientist assumes that the relationship between diabetes and certain presumed risk factors (e.g., age, weight, diet, etc.) follows a multivariate logistic regression, or when a statistician assumes that the test statistic computed from a clinical trial follows a t-distribution with a certain degrees of freedom. Initially, the probabilities emerging from these analyses are *academic* (as in ‘not of practical relevance, of only theoretical interest’) because they attain their meaning by *fiat* (an arbitrary degree or pronouncement) instead of by a concrete connection to the real world. And because these probabilities are academic, so are the conclusions those probabilities support.

Conclusions drawn from academic analyses become *real* only when the hypothetical assumptions on which they are based become tied to reality.

> *Real probabilities are backed by something real and are about something real.*

For a probability to be 'backed by something real' it must have a tangible guarantor, i.e., something must serve as collateral to ensure that the purveyor of the probability ('the probabilist') is acting in good faith. For a probability to be 'about something real' the statement that the probability is about must be decidable, in that there must be clear criteria to determine whether or not the object of the probability statement did or did not happen, is or is not true.

With this understanding, probabilities used in stock option pricing are academic *until an option is bought or sold on the basis of those probabilities.* Once the transaction is made, the probability becomes testable through the realized profit or loss of that transaction. In the same way, probabilities and derivatives of probabilities such as P-values, confidence intervals, etc. reported in scientific literature are academic *unless and until those probabilities can be tested for their accuracy.* (N.B. Although realizing a loss on any given transaction does not indicate that the associated probability calculation is wrong (and similarly realizing a gain on any given transaction does not indicate that the associated probability calculation is right), sustained losses (gains) over a long series of transactions provide an objective way to assess systematic error (or soundness) in these calculations.)

The Fundamental Principle of Probability. The distinction between academic and real probabilities gives rise to what I call the *Fundamental Principle of Probability* (FPP), which establishes the connection between academic and real probabilities by tying academic probabilities to real outcomes. The FPP is a common sense concept, intuitively understood by almost every stock trader, bookie, and man on the street:

> *If you assign a probability to an outcome happening, then you must be willing to accept a bet offered on the other side (that the outcome will not happen) at the correct implied odds.*

By the FPP, when you claim that the probability of 'A' is p, then you are implicitly offering odds of $p/(1-p)$ for a bet on 'not-A'. If I choose to take you up on your offer, then you win whatever money I risked if A happens, and I win p of your dollars for every $1-p$ dollars I risked if A does not happen.[2]

According to the FPP, real probabilities consist of three ingredients:

(I) A decidable statement A to which the probability statement applies.

(II) An exposure limit, i.e., the maximum amount of money that one is willing to risk on the probability assessment of A.

[2] Technical note: this interpretation of probabilities as prices of bets is fundamental to the 'radical subjective Bayesian' (RSB) philosophy of de Finetti, Savage, and others. The difference between the RSB, as a philosophical stance, and the FPP, as a practical principle, is the same difference between academia and reality. RSB probabilities are merely hypothetical degrees of belief with only an academic connection to betting, while real probabilities are actual betting quotients, as realized by the real consequences of stating those probabilities (the potential for financial loss or gain).

(III) A probability p that the assertion A turns out to be true.

Academic probabilities often fail to satisfy (I) and almost always fail to satisfy (II). In current academic science, the exposure limit in (II) is implicitly set to 0, which according to the FPP conveys no confidence in the probability assessment, and therefore no credibility to its associated conclusion. To make scientific conclusions more credible, scientists need to up the ante on their claims. It may well be that the scientist is convinced by this own analysis and provides a compelling argument in its favor. But the probabilities backed up by these words are meaningless unless backed up by the FPP.

Technically, the probability p in (III) is a 'lower probability' for A, in the sense that it determines the odds $p/(1-p)$ of a bet *against* A. (Note that it is not assumed that p is the price of a 'fair bet', as offering fair bets would expose the probabilist to potentially ruinous fluctuations in wealth without any potential long-run gain. The purpose of the FPP is to make probabilities meaningful, not to force scientists to gamble or expose themselves to ruin.) If p is a legitimate lower bound on the probability of A, then the implied odds $p/(1-p)$ for a bet against A will be favorable to the probabilist in the long run. If p is too optimistic, however, the implied odds will overpay for losses, causing the probabilist to suffer financial loss in the long run and providing evidence against the initial probability assessment.[3]

Academic probabilities and the replication crisis. At its core, the replication crisis is a statistical and probabilistic crisis, caused by a combination of misapplied statistical methods, misinterpreted probabilities, and unchecked misconduct. The crisis is allowed to persist by confusing academic probabilities, and the conclusions they are claimed to support, for something real. The FPP provides a built-in mechanism that forces users of statistical and probabilistic methods to report more meaningful probabilities (or else report no probabilities at all), without any need for validation by editors, peer reviewers, or other so-called experts. It's a simple premise: instead of enacting further bureaucratic regulations in hope of resolving the replication crisis, simply force scientists to back up their claims according to the FPP.

To appreciate how implementing the FPP would affect the reporting and evaluation of scientific conclusions, first consider the role it plays in the casino industry. With

[3] Technical note: The FPP gives a probability statement real meaning by forcing the purveyor of probability to act as the 'house' for anyone wishing to test the probability assessment. The probabilist may also wish to state a probability for the contrapositive of A (i.e., that A does not happen), written $-A$. In this case, the lower probability p' for $-A$ offers an implied odds of $p'/(1-p')$ for a bet on A. Together the probabilities p and p' give a range for the actual probability of A:

$$
\begin{aligned}
p &\leq P(A) \leq 1-p' \\
p' &\leq P(-A) \leq 1-p.
\end{aligned}
$$

If these are not proper intervals, meaning that $p+p' > 1$, then the probabilist exposes himself to sure loss, as I can bet $1-p$ on A at odds of $p/(1-p)$ and $1-p'$ on $-A$ at odds of $p'/(1-p')$. If A happens, then I win $(1-p) \times p/(1-p) = p$ for my bet on A and lose $1-p'$ for my bet on $-A$, for a total gain of $p-(1-p') = p+p'-1 > 0$. If $-A$ happens, then I win $(1-p') \times p'/(1-p') = p'$ for my bet on $-A$ and lose $1-p$ for my bet on A, for a total gain of $p'-(1-p) = p+p'-1 > 0$.

the exception of professional advantage players,[4] it is well known that casino games are unfavorable to the gambler. Importantly, this knowledge comes from common sense, and not from a formal statistical analysis of casino games (an analysis which the average person is ill-equipped to carry out). It is obvious that the odds offered by any major casino must favor the house by the simple facts that (1) the casino willingly accepts any bet at its stated odds and (2) casinos are profitable businesses.

With respect to the specific odds offered at specific casinos, one can ask how I know that a 7 is rolled with less than 20% frequency at the craps tables at Australia's Crown Melbourne Casino. I claim to know this even though I've never visited the Crown Melbourne, much less inspected its dice, tables, or dealers to confirm that the game is (more or less) fair. And as emphasized above, I'm not basing my knowledge on a theoretical STAT 101 calculation—under the standard model of two 'fair dice' the probability of rolling a 7 is 16.7%—since I have no way to assess the assumptions behind that calculation. In fact, the exact probability (i.e., frequency) of rolling a 7 at the Crown Melbourne very well might be 16.7%, or 18%, or 14%. I don't know, and I don't care. No matter what the actual frequency is, I'm pretty sure that it must be less than 20%. And I know this because the Crown Melbourne offers odds of 4-to-1 (an implied probability of $1/(4+1) = 0.20$) for a one roll bet on 7 *and the Crown Melbourne is still in business*. Whether the true frequency is 16.7%, 18%, or 14%, as long as it is lower than 20%, the casino will win in the long run, and the gambler will lose.

Because of the survival mechanism built in to the FPP—casinos that state bad probabilities either revise their odds or go bankrupt—and the observation that the Crown Melbourne casino is still surviving (in fact, thriving), I can be reasonably sure that the probability bound implied by its odds is legitimate without understanding how the Crowne Melbourne determines its odds and without even knowing the rules of craps. I know this all because the casino's probabilities are real: if the implied odds were too high, the casino would go broke; if they were too low, then the casino would have too few customers.

If we don't need inside information to determine that the casino's odds give valid probability bounds for the outcomes of its games, why do we need editors, associate editors, and referees to evaluate the soundness of methods that scientists use to compute P-values, confidence intervals, Bayes factors, and other probabilistic measures of evidence? Just like casinos don't offer overly generous odds to attract more customers (because they would quickly go bankrupt), wouldn't it be nice if scientists didn't even try to publish hacked or overly optimistic P-values or other probabilities, because doing so would put them out of the scientific business? It's a simple idea with a simple implementation thanks to the Fundamental Principle of Probability.

[4] 'Advantage players' are gamblers who bet with the odds in their favor. Perhaps the best known example of advantage play is card counting in blackjack [Tho66]. More recently, Phil Ivey and Cheung Yin Sun won more than $20 million dollars using a technique called edge sorting to get an advantage while playing baccarat [Wik]. Other examples of advantage play in sports betting, video poker, roulette, craps, and other 'unbeatable' casino games are discussed in Bob Dancer and Richard Munchkin's 'Gambling with an Edge' podcast [DM].

Replication and the Fundamental Principle of Probability. In scientific terms, every roll of the dice, spin of the roulette wheel, and hand of blackjack is a test of the casino's stated odds, and thus also the probabilities implied by those odds. Every bet against the house is an attempt to falsify the casino's implicit probability claims. And the fact that casinos consistently profit from these games, and gamblers consistently lose, can (and should) be interpreted as compelling evidence that the odds are set correctly.[5]

Like casino odds, bureaucratic assessment of scientific methods would become unnecessary if scientific conclusions were also backed by *real* probabilities. Overly optimistic probabilities about the reliability (i.e., replicability) of a given finding would incentivize replication attempts; and overly conservative probabilities would understate the importance and reliability of a given conclusion. Accordingly, the Fundamental Principle of Probability suggests a straightforward and objective way to report conclusions and carry out replication attempts in the scientific literature by laying out the following conditions when reporting their conclusions.

> **Replication Protocol 1.** Describe the protocol under which a replication attempt should proceed.
> **Replication Criteria 1.** State the criteria on which the results of the protocol will be considered a successful replication.
> **Replication Probability 1.** State the probability that this procedure will result in successful replication.
> **Replication Exposure 1.** State an amount of money that the scientists behind these claims are willing to expose for tests of their conclusions, in accordance with item (II) of the FPP above, and put this amount of money in escrow.
>
> Continue with additional Replication Protocols, Criteria, and Probabilities 2, 3, 4, etc. as desired.

The stated protocols, criteria, and probabilities provide a concrete basis on which to test findings according to the FPP. This in turn constrains the stated replication probability by forcing the authors, their institutions (if any), the journal publishing the work (if any), and any editors, associate editors, and referees involved in the publication decision to back up their claims with something real.[6]

[5] By 'correctly' here I mean that the probabilities implied by those odds give a valid lower bound on the true probability (i.e., frequency) of occurrence.

[6] Depending on the field, this collateral could be $500, $5,000, $50,000, or $500,000 per claim. The appropriate amount of 'skin in the game' should be roughly proportional to the cost of carrying out the research and what the involved parties stand to gain from the publication. The point is that this collateral should be sufficiently large that the authors should not be willing to lose the money in exchange for publication. So if it cost $1 million (most or all of which is usually taxpayer money) to perform the research, then it is reasonable to post an additional $500K to back up the claims of that research. If the research costs only $1K, then $500 should be sufficient. As a practical matter, at least part of the cost of the replication attempt, performed by an agreed upon third party, should be defrayed by this collateral.

After publication, the collateral is locked in escrow for a fixed period of scrutiny (say 2 years) during which other scientists can choose to bet against replication at the odds implied by the stated replication probability.[7] A scientific team (a 'skeptic') wishing to test the claimed replication probability will put the amount they are willing to risk in escrow while the replication attempt is carried out. In doing so, the skeptics agree to the terms stated in the original authors' replication protocol and criteria, which would set in motion a replication attempt by an agreed upon neutral third party.[8] The outcome of that replication attempt determines the outcome of the bet between the scientists making the claims (the probabilists) and those testing the claims (the skeptics).

This process continues for the pre-specified period of scrutiny or until the escrow money runs out. If the escrow money runs out, then the authors either post more collateral or the result is archived with clear documentation of the outcome of the replication attempts along with a record of the author's net monetary gain/loss. Authors who nevertheless stand by their claims could post more escrow money to encourage additional testing of their results, in hope of validating their initial findings. After the period of scrutiny is over, regardless of the outcome the results of the replication attempts are published along with the final tally of gains/losses at the stated replication probabilities. It is important to note that while financial gains/losses throughout the replication process do not necessarily indicate that the initial claims are right/wrong, they can be reasonably interpreted as evidence for/against the initial claims, and such evidence should be documented in the published scientific record.

The authors are free to revise their probabilities (upward or downward) at any point during the period of scrutiny, and the odds at which the replication studies are offered will be revised accordingly. For example, if the initially stated replication probability of 80% is too conservative (e.g., the true replication probability should be 90%), then there may be few attempts to replicate the initial claims, because the odds against replication are unfavorable. In response to this, the authors may wish to increase their replication probability to 85%, serving two beneficial purposes from the standpoint of the original authors. First, it strengthens the original claims—because the probabilities stated by the authors are *real* in the sense of being backed by the FPP, the higher replication probability makes the original claims stronger. Second, the higher probability of replication offers a better price to anyone wishing to challenge the results (5.6-to-1 instead of 4-to-1), and so could entice more activity in trying to replicate/falsify the original results. In turn, successful replication makes the claims

[7] Also stated in the article, or as part of the journal's replication policy, should be a standard protocol for how replication attempts will be adjudicated. Importantly, the replication attempt must occur after the skeptic has decided to bet against replication and has placed his/her money in escrow. This replication attempt should be carried out by a trusted, neutral third-party team of scientists, for which the costs of carrying out the replication are funded in whole or in part by the initial scientific team and the journal publishing the claim.

[8] The precise protocol by which such third parties will be chosen and the truthfulness of the conclusions assured is a practical issue of much the same flavor as ensuring that the current peer review process is properly administered. I don't discuss this further here.

in the article more credible. If the authors' assessment of the true probability at 90% is correct, these replication attempts should generate additional research funds in the long term. Furthermore, successful attempts to replicate will make the authors' findings more credible.

Alternatively, the authors could adjust their replication probabilities downward, say from 80% to 70%, if they determine that their initial probabilities were overly optimistic. Either way, if the stated conclusions are reliable, and the probabilities are well-calibrated, then the scientists, their universities, and the journals behind good and precise science will be rewarded, both financially and academically, for drawing sound conclusions. If the conclusions are unreliable, then these scientists, the universities who support them, and the journals who publish bad science will be penalized.

What will come of this? The above process leverages the built-in correction mechanism of statistical methods (through the FPP) to improve the reliability of scientific claims based on statistical and probabilistic methods: authors should not state overly conservative probabilities, or else there will be little interest in their conclusions and few attempts to replicate their findings; but they should also not state overly optimistic probabilities, or else they are offering too good of a price for replication attempts that are likely to cost them money in the long run. Such financial loss has the direct impact of depriving researchers who publish overly optimistic replication probabilities of research funding. Also, because these outcomes are documented in the final published version of the results, the failed replication attempts may also lead to a loss in credibility of the researchers involved in the failed claims. Conversely, a gain of funds from the replication process will have the opposite (positive) effect on the reputation and financial status of the scientists behind the initial claims. Most important of all, the scientific literature should become more accurate and more complete as a result.

Stated replication probabilities should be more conservative. Assuming its dice are fair, the casino knows that the true odds of rolling 7 are 5-to-1 (implied probability of 16.7%), but it offers only 4-to-1 (implied probability of 20%). In terms of the replication probabilities above, the casino is claiming that the probability of *not* rolling a 7 is 80% even though the theoretical probability is 83.3%. A casino that overstates the probability, say, at 85% (for 5.6-to-1 implied odds), would suffer financial loss as a result.

By analogy, the scientist who uses a statistical method to compute the probability of replication at 90% is incentivized to quote a more conservative probability, say 80% or 85%, in publication. These understated probabilities account for uncertainty due to assumptions of the model, approximations, and other factors beyond the scientists' control. Scientists who publish a higher figure to convey confidence in the outcome—the academic equivalent of a bluff—risks losing the money posted as collateral when their bluff is called. This in turn will create a negative impression of their stated conclusions and deprive these scientists of resources for future research. If, however, the probability is too conservative (say, 25% instead of 80%), then the claim will be too weak to generate sufficient interest in replication, which in turn undermines the significance of the reported finding.

Probabilities reported in the literature will be more meaningful. Even ignoring the divide between academia and reality, the specific probabilities that appear in scientific writing are often interpreted qualitatively. Without the opportunity to test probabilities or model assumptions (according to the FPP), there is little conceptual difference between a probability of 90% or 95%. Both probabilities are high, and would usually be interpreted as evidence in support of a given claim. But when bound by the FPP there is a big difference between 90% (9-to-1 odds) and 95% (19-to-1 odds). If the true probability is somewhere in between, say 92.5%, then stating 90% instead of 95% is the difference between a positive or negative return on the funds put in escrow, giving the reported numerical probabilities a more precise meaning.

Increased replication attempts. Rewarding scientists for correctly identifying overly optimistic published claims provides an incentive to spearhead replication attempts. Aside from gaining better clarity about the published literature, researchers who call for the replication of flawed claims will be rewarded with additional funds for their own research. Contrast this to the current situation in which a number of people bemoan the fact that replication attempts are too few because there are no incentives (financial or otherwise) for replicating the results of other scientists.

No protection for 'normal science'. Because of the way scientific results are currently evaluated, by the gatekeepers in charge of peer-reviewed journals, there is a tendency for findings that go against the *status quo* to be suppressed [Kuh12]. Shifting from academic to real probabilities, and therefore academic to real science, would eliminate the need to suppress anything that is submitted for publication in the scientific literature. The scientific establishment can easily refute outrageous and bogus claims on objective grounds by betting against replication at the implied odds. If the establishment is correct, then the scientists behind the inaccurate claims will suffer financial loss, and the establishment a financial gain; and conversely if the establishment is incorrect. In this way, controversial results that go against the scientific paradigm need not be filtered out of the literature, as is currently the norm, but rather can be put under scrutiny in a transparent and objective manner, further strengthening the predominant paradigm when these new results are refuted and expediting progress when the controversial results stand up to scrutiny.

Investing in better science. Some critics will cry foul that science isn't a betting game and that the financial incentives of the above proposal will corrupt science. But while I've used the language of betting here for illustration, the proposal is better thought of as a mechanism for investing in good science/scientists and divesting from bad science/scientists, with the intention to improve the overall quality and reliability of published science. Such a result would clearly improve, not corrupt, the current state of science. The way it works right now is that scientists report fiat probabilities, and implicitly state odds, with few consequences for passing bad probabilities, fake probabilities, counterfeit probabilities, or fraudulent probabilities. The result is the replication crisis.

For good scientists, good journals, good editors, and good referees, the proposal here is an investment in good science. Better scientists will be rewarded with more resources to do more, better science. Bad scientists will be penalized by losing resources, so that they must do less science, or else re-focus their efforts on quality over quantity. Similarly, universities that employ good scientists and journals that consistently publish reliable science will be rewarded. Those that support bad science from bad scientists will be penalized. Instead of letting universities skim millions of dollars of taxpayer money as 'indirect costs' from grants, funding agencies should mandate that some or all of those indirect costs first be invested in the science conducted under that grant, as part of the escrow. Whatever share of that escrow money is left after the period of scrutiny goes to the university in indirect costs. If the escrow money is lost, then the university gets no overhead. The indirect costs instead go to the institutions of the scientists who correctly invested in the attempt to falsify the bad conclusions. As a result, taxpayers who fund the research will get a better return on their investment because their tax money will be automatically allocated to more prudent, honest researchers based on results, instead of on the subjective determination of a grant panel, editorial board, or anonymous referee.

And as far as incentives are concerned: are the incentives of prestige, tenure, promotion, and money not already a factor in how scientists behave? Aren't the arbitrary ways in which these awards are conferred at least partly to blame for rampant P-hacking, fraud, and other questionable research practices in the scientific literature? In fact, these sociological factors are often cited as the catalyst for the 'results-oriented' attitude that spawns the replication crisis, and which the proposal here seeks to correct. Even with the FPP, the same incentives will be present, but the FPP offers improvement by making it harder to achieve these incentives by asserting spurious, careless, or fake probabilities.

Reforming science with real probability. When Disraeli spoke of "lies, damn lies, and statistics" he was referring to statistics in the absence of the FPP, or to probability without "skin in the game" [Tal18]. The suggestion here to merge probabilities reported in the scientific literature with the FPP is a proposal to improve the reliability of the scientific literature by giving real meaning to reported probabilities, and thus real meaning to the conclusions based on them, all without imposing any extra requirements, administrative work, or other restrictions on the way research is conducted. Rather than being burdened by additional constraints, scientists will be liberated to run studies as they deem appropriate, without the need to apply a specific method in order to appease a referee or be published in a certain journal. With this freedom comes accountability for reported results, as ultimately the stated replication probability will have to stand on its own merits when tested according to the scientists' prescribed replication protocol. This system is designed so that good scientists succeed, bad scientists fail, and conclusions reported in the literature are more reliable on the whole.

When implemented in this way, the FPP might be thought of as the Hippocratic Oath of probability, an oath not to expose others to the potential harm of statistical methods without exposing oneself to that same downside risks. Those uncomfortable taking the risks associated with their methods shouldn't be evaluating their results using probabilities, just like those who can't stand the sight of blood aren't cut out for medicine and athletes with a weak chin aren't cut out to be boxers.

References

[Cha14] C. Chambers, *Registered Reports: A step change in scientific publishing*, Accessed at https://www.elsevier.com/reviewers-update/story/innovation-in-publishing/registered-reports-a-step-change-in-scientific-publishing on May 30, 2018 (2014).

[Col15] Open Science Collaboration, *Estimating the reproducibility of psychological science*, Science **349** (2015), no. 6251.

[DM] B. Dancer and R. Munchkin, *Gambling with an Edge Podcast*, Accessed at https://www.lasvegasadvisor.com/gambling-with-an-edge/ on June 20, 2018.

[Ioa05] J.P.A. Ioannidis, *Why Most Published Research Findings Are False*, PLoS Medicine **2** (2005), no. 8, e124.

[Kuh12] Thomas S. Kuhn, *The Structure of Scientific Revolutions, 4th edition*, University of Chicago Press, 2012.

[Tal18] N.N. Taleb, *Skin in the Game*, 2018.

[Tho66] E.O. Thorp, *Beat the Dealer: A Winning Strategy for the Game of Twenty-One*, Vintage, 1966.

[Wik] Wikipedia, *Edge sorting*, Accessed at https://en.wikipedia.org/wiki/Edge_sorting.

Department of Statistics & Biostatistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA

E-mail address: hcrane@stat.rutgers.edu