Replication Crisis, Prediction Markets and Fundamental Principle of Probability

Harry Crane

Department of Statistics
Rutgers

October 22, 2018

Key Ideas

- **Goodhart's Law**: When a measure becomes a target it ceases to be a good measure.
  - *Theory*: P-value measures the amount of evidence data provides for a scientific hypothesis.
  - *Practice*: P-value < 0.05 is the target that determines whether results are published or not.

- **Cournot's Principle**: an event of small or zero probability singled out in advance will not happen. (Shafer)
  - Gives probabilities empirical meaning.

- **Fundamental Principle of Probability**: If you assign a probability to an outcome, then you must accept a bet offered on the other side at the correct implied odds.
  - Gives probability statements both credibility and meaning.
  - *Meaning*: Probability p of A means an offer to bet against A at $p/(1-p)$ odds.
  - *Credibility*: Someone claiming a probability p of A must believe the probability is *at least* p or else risk long-term loss.

H. Crane. (2018). The Fundamental Principle of Probability. https://www.researchers.one/article/2018-08-16


Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Why?

1. **Low Power**: The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
2. **Multiple Testing**: The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
3. **Researcher Degrees of Freedom**: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
4. **Incentives**: The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.

“Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias” (Ioannidis)

What proportion of results should replicate?

	H_0 true	H_0 false
Proportion	ϕ	$(1-\phi)$
Reject	$\alpha\phi$	$(1-\beta)(1-\phi)$
Not Reject	$(1-\alpha)\phi$	$\beta(1-\phi)$

Table: α is Type-I error rate; β is Type-II error rate; $(1-\phi)/\phi$ is prior odds.

- **Type-I error rate**: $\Pr(P < \alpha \mid H_0 \text{ true}) = \alpha$.
- **Type-II error rate**: $\Pr(P > \alpha \mid H_0 \text{ false}) = \beta$.
- **Prior odds** $(1-\phi)/\phi$: ratio of false to true H_0 among all those tested.
- **False positive rate** (FPR): proportion of false positives among all $P < \alpha$

$$\text{FPR}(\alpha, \beta, \phi) = \frac{\alpha\phi}{\alpha\phi + (1-\beta)(1-\phi)}. \tag{1}$$

- **Replication rate** (RR): $\text{RR}(\alpha, \beta, \phi) = 1 - \text{FPR}(\alpha, \beta, \phi)$.

Empirical Evidence

Large-scale replication studies have successfully replicated

- 39% of findings (37 out of 97) in psychology.[1]
- 61% (11 out of 18) in experimental economics.[2]
- 62% (13 out of 21) of social science articles published in *Science* and *Nature*.[3]

Does this contradict theoretical expectations?

α	β	$\phi/(1-\phi)$	FPR	RR
0.05	0.20	1:1	6%	94%
0.05	0.20	1:10	36%	64%
0.005	0.20	1:1	$<$ 1%	$>$ 99%
0.005	0.20	1:10	5%	95%

- Maybe 60% replication rate shouldn't be surprising – reflects testing parameters (α, β, ϕ).
- But published findings are/should be based on more than single hypothesis test – Theoretical FPR should be upper bound.

[1] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 2015.

[2] C. Camerer et al. Evaluating replicability of laboratory experiments in economics. *Science*, 2016.

[3] C. Camerer et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2, 637–644, 2018.

What to do?

1. "Redefine Statistical Significance"
   - Change default cutoff for statistical significance from 0.05 to 0.005.
   - Why? FPR under 5% level is too high even in absence of misconduct or reporting issues.
2. Pre-registration and results-blind peer review
   - Decide whether to publish results based on methods, not results.
   - Why? Incentivize sound methods over results.
3. Prediction Markets
   - Use betting markets to make better publication decisions.
   - Why? Incentivize more accurate assessment of replication chances by scientific community prior to publication.
4. Fundamental Principle of Probability
   - Authors set their own replication criteria and probability and back up claims with their own money.
   - Why? Incentivize more accurate and honest assessment of replication by authors. Penalize inaccurate and dishonest claims (i.e., P-hacking).

Observation: Current paradigm promotes skewed incentives, reporting bias, and replication crisis.

Solution: Change the paradigm to eliminate reporting bias (publish everything), align incentives (reward accuracy, punish inaccuracy), and directly target replication.

Title: Redefine Statistical Significance

Authors: Daniel J. Benjamin[1]*, James O. Berger[2], Magnus Johannesson[3]*, Brian A. Nosek[4,5], E.-J. Wagenmakers[6], Richard Berk[7, 10], Kenneth A. Bollen[8], Björn Brembs[9], Lawrence Brown[10], Colin Camerer[11], David Cesarini[12, 13], Christopher D. Chambers[14], Merlise Clyde[2], Thomas D. Cook[15,16], Paul De Boeck[17], Zoltan Dienes[18], Anna Dreber[3], Kenny Easwaran[19], Charles Efferson[20], Ernst Fehr[21], Fiona Fidler[22], Andy P. Field[18], Malcolm Forster[23], Edward I. George[10], Richard Gonzalez[24], Steven Goodman[25], Edwin Green[26], Donald P. Green[27], Anthony Greenwald[28], Jarrod D. Hadfield[29], Larry V. Hedges[30], Leonhard Held[31], Teck Hua Ho[32], Herbert Hoijtink[33], James Holland Jones[39,40], Daniel J. Hruschka[34], Kosuke Imai[35], Guido Imbens[36], John P.A. Ioannidis[37], Minjeong Jeon[38], Michael Kirchler[41], David Laibson[42], John List[43], Roderick Little[44], Arthur Lupia[45], Edouard Machery[46], Scott E. Maxwell[47], Michael McCarthy[48], Don Moore[49], Stephen L. Morgan[50], Marcus Munafò[51, 52], Shinichi Nakagawa[53], Brendan Nyhan[54], Timothy H. Parker[55], Luis Pericchi[56], Marco Perugini[57], Jeff Rouder[58], Judith Rousseau[59], Victoria Savalei[60], Felix D. Schönbrodt[61], Thomas Sellke[62], Betsy Sinclair[63], Dustin Tingley[64], Trisha Van Zandt[65], Simine Vazire[66], Duncan J. Watts[67], Christopher Winship[68], Robert L. Wolpert[2], Yu Xie[69], Cristobal Young[70], Jonathan Zinman[71], Valen E. Johnson[72]*

One Sentence Summary: We propose to change the default *P*-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

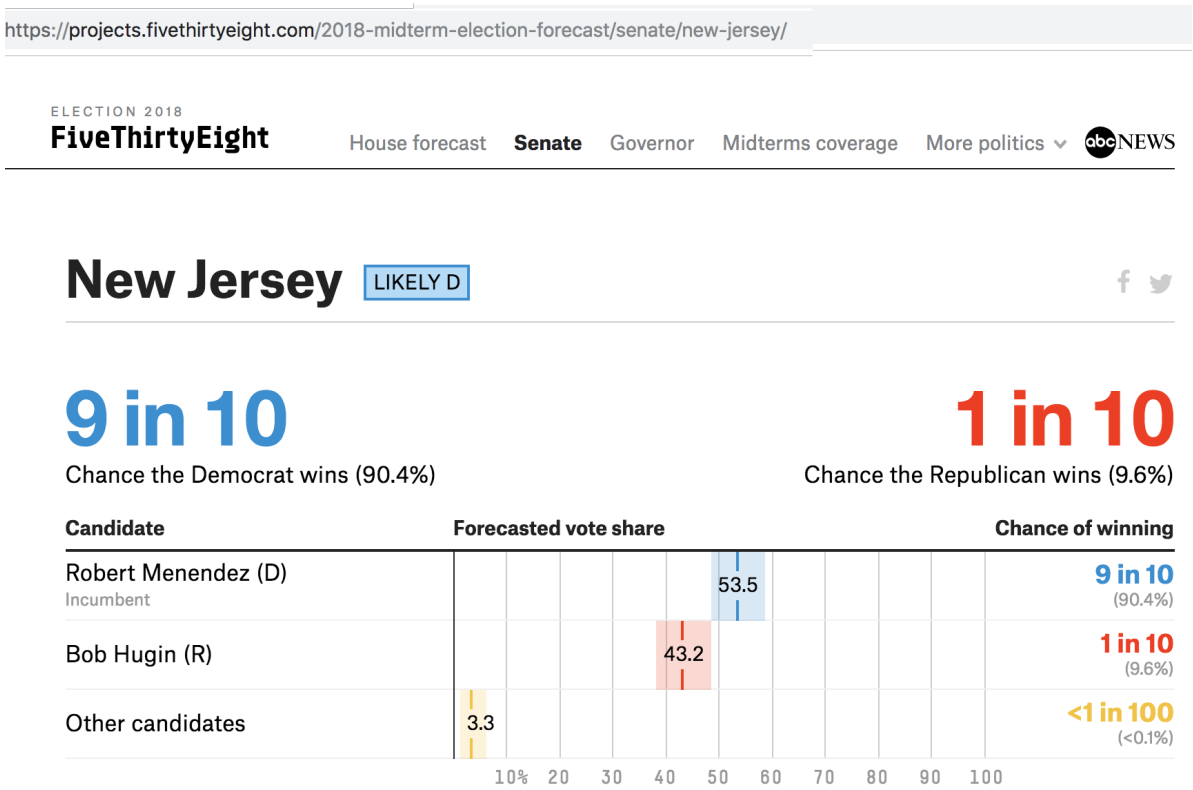
Summary of RSS argument

- Benjamin, et al (2017, *Nature Human Behavior*) propose to "redefine statistical significance" from $P < 0.05$ to $P < 0.005$.
- Claimed benefits of RSS:
  - Reproducibility would "immediately improve".
  - False positive rates will fall as low as 5% and improve by factors greater than 2.
  - Replication rate will approximately double.
- Other claims:
  - Lower cutoff supported by "critical mass of researchers".
  - 0.05 cutoff is "leading cause of non-reproducibility".

- **Flaws in argument**:
  - Numerous — see Amrhein–Greenland; Lakens, et al; Trafimow, et al; Crane.
- **Main criticism**:
  - Argument based on hidden/misleading assumptions.
  - Leads to exaggerated conclusions.
- **Irony**:
  - RSS article in *Nature Human Behavior* ignores an obvious part of human behavior (i.e. P-hacking and incentives of publication).
  - Makes non-replicable claims in argument claiming to improve replication crisis.

H. Crane. (2018). The impact of P-hacking on "Redefine Statistical Significance". Basic and Applied Social Psychology.

“Expert” Opinion



- **FiveThirtyEight Estimate**:

$$\Pr(\text{Menendez re-elected}) = 0.904$$
$$\Pr(\text{Menendez not re-elected}) = 0.096$$

Prediction Markets (from PredictIt.com)

Will Bob Menendez be re-elected to the U.S. Senate in New Jersey in 2018?

Latest Price: 78¢ ↓ 1¢

Buy Yes — Click to match Offers starting at **78¢**, or to make your own, lower Offer.

Buy No — Click to match Offers starting at **23¢**, or to make your own, lower Offer.

If this prediction comes true, PredictIt will redeem all Yes shares at $1. Shares in No will have zero value. If this prediction does not come true, PredictIt will redeem all No shares at $1. Shares in Yes will have zero value.

Data | **Rules** | Prices

Bob Menendez shall be the winner of the 2018 general election for U.S. Senator from New Jersey. PredictIt may determine how and when to settle the market based on all information available to PredictIt at the relevant time.

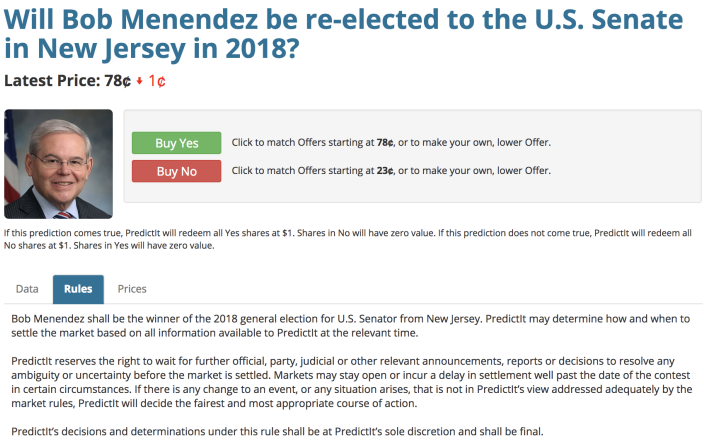
PredictIt reserves the right to wait for further official, party, judicial or other relevant announcements, reports or decisions to resolve any ambiguity or uncertainty before the market is settled. Markets may stay open or incur a delay in settlement well past the date of the contest in certain circumstances. If there is any change to an event, or any situation arises, that is not in PredictIt's view addressed adequately by the market rules, PredictIt will decide the fairest and most appropriate course of action.

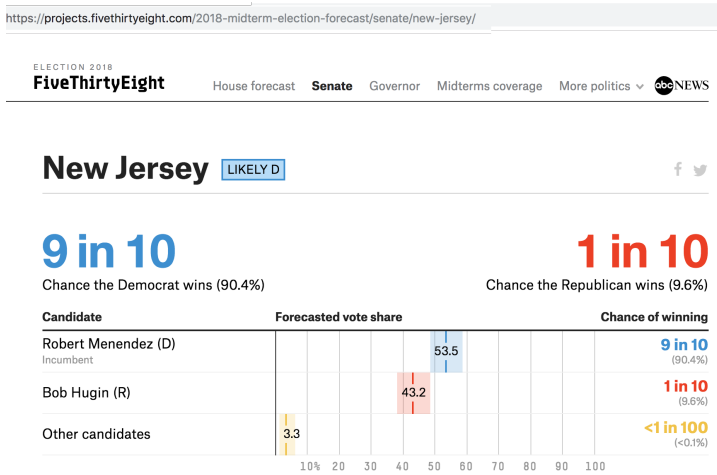
PredictIt's decisions and determinations under this rule shall be at PredictIt's sole discretion and shall be final.

- **Prediction Market Estimate**:

$$\Pr(\text{Menendez re-elected}) = 0.78$$
$$\Pr(\text{Menendez not re-elected}) = 0.23$$

The Experts or the Masses?





Prediction Market: $\Pr(\text{Menendez re-elected}) = 0.78$

538 Forecast: $\Pr(\text{Menendez re-elected}) = 0.90$

Which is more accurate?

- If 538 estimate is right, the market is offering a 12% edge to bet on Menendez.
- Either (i) markets very inefficient or (ii) 538 estimates are unreliable. (Maybe both.)

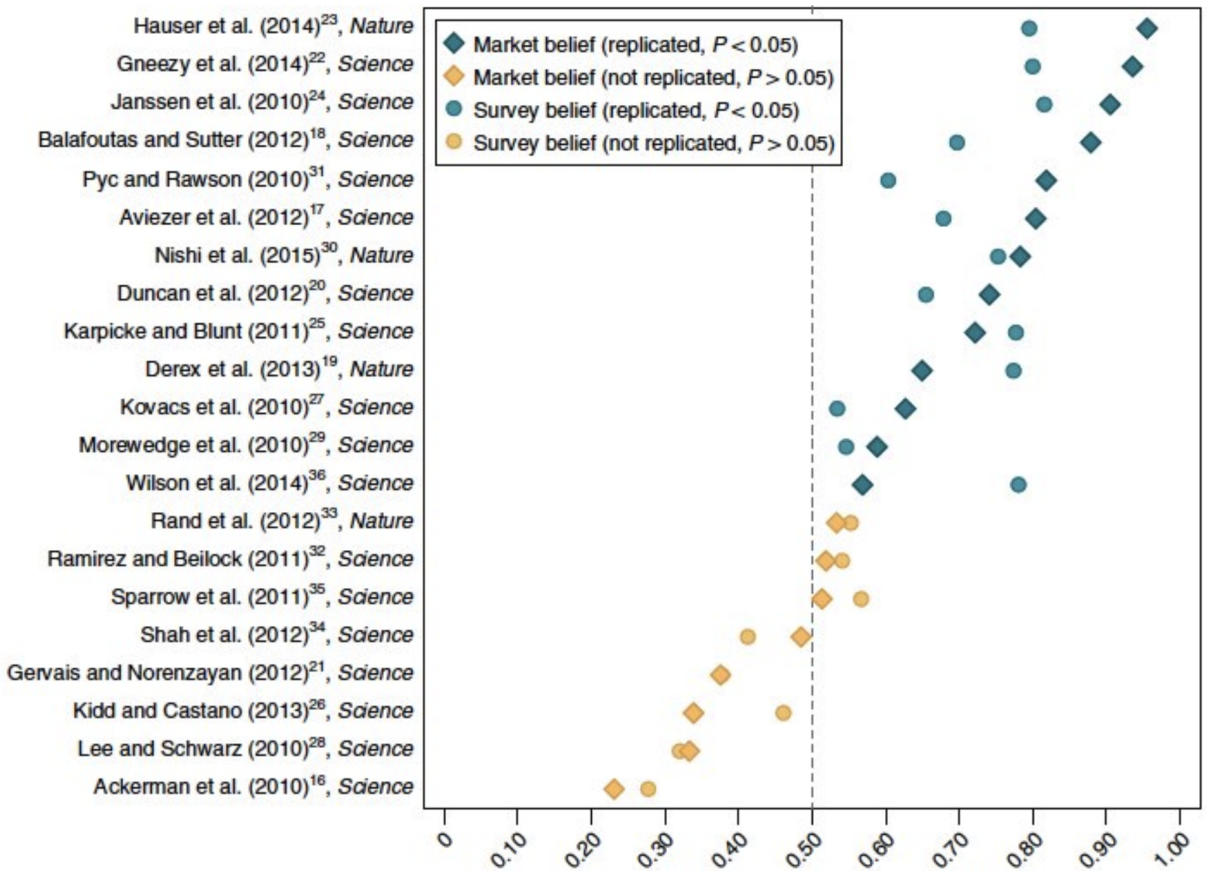
Who has more incentive to be accurate?

- Who is more likely to suffer from slight deviations from accuracy? Markets.

Prediction Markets and Replication

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16*], Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]



Could Gambling Save Science?

Idea: Use Prediction Markets to get more accurate measure of scientific community's assessment of replication.

1. Submit paper to journal $\rightarrow$ betting market opens up.

   Conditional on a replication attempt, the claimed results will replicate:

   YES: *$0.67*
   NO: *$0.33*

   - If replication attempt successful, **YES** worth \$1.00, **NO** worth \$0.00.
   - If replication unsuccessful, **YES** worth \$0.00, **NO** worth \$1.00.
   - If no replication attempt within (say) 1 year, then no action.

2. After 2 months (or so), journal decides whether to accept/reject paper. Can use prediction market price in its decision.
3. If accepted, the prediction market price after 2 months is published with article.

R. Hanson. Could Gambling Save Science? Encouraging an Honest Consensus.
R. Hanson. Shall We Vote on Values, But Bet on Beliefs?

Pros and Cons of Replication Prediction Markets

Recall Ioannidis's reasons for low replicability:

1. **Low Power**: Can and should be priced in by the prediction market.
2. **Multiple Testing**: Can be priced in by bettors, but information asymmetry between authors and market.
3. **Researcher Degrees of Freedom**: Limited because authors specify replication criteria prior to submission.
4. **Incentives**: Author incentive to pump up the price of replication.
5. **"Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias"**: High prediction market prices may just be more accurate measure of prevailing bias.

New boss same as the old boss?

Main issue: Authors still benefit from information asymmetry.

- Authors benefit from inaccurate high market prices $\Rightarrow$ incentive to mislead.
- Authors punished by inaccurate low market prices $\Rightarrow$ paradigm-shifting and unpopular ideas suppressed.
- P-hacking, QRPs, etc. still viable career strategy.
- Market probability is a more accurate reflection of the community opinion, but can't account for unreported insider (author) information and authors still have incentive to withhold information.

So what do we expect to change?

Fundamental Principle of Probability and Replication

Fundamental Principle of Probability (FPP): If you assign a probability to an outcome, then you must accept a bet on the other side at the correct implied odds.

Idea: Apply FPP to get more accurate measure of author's assessment of claims and directly tie accuracy of author assessment to outcomes of replication.

1. Publish paper along with:
   - (I) **Replication Criteria**: run [experimental protocol] for sample size $N \geq 100$, compute test statistic T. If $|T| > t_c$, then declare successful replication; otherwise, not.
   - (II) **Replication Probability**: the above procedure will replicate with probability p.
   - (III) **Exposure Limit**: Authors put some amount of money (e.g., \$10,000) in escrow for pre-determined period of scrutiny (say, 2 years).
2. During 2 year period, anyone else can put up $\$A$ to gain $\$p/(1-p) \times A$ in event of failed replication.
   - If replication attempt successful, Authors gain $\$A$.
   - If replication unsuccessful, Authors lose $\$A \times p/(1-p)$.
   - If no replication attempt, then no action.
3. Results of replication attempts (if any) published as addendum to original article.
4. All papers (even failed replications) are published.

Pros and Cons of FPP

Recall Ioannidis's reasons for low replicability:

1. **Low Power**
2. **Multiple Testing**
3. **Researcher Degrees of Freedom**

   All 3 can and should be priced in by authors. Failure to do so directly exposes authors to loss.

4. **Incentives**: Author incentive to deflate the probability of replication. Think about how casinos set odds.
5. **"Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias"**: Claimed probabilities reflect only the authors bias/naivety/ misguidedness/expertise. If authors are wrong, opportunity for other researchers to gain. If community is wrong, authors gain.

Benefits of FPP

- **Incentivize accurate reporting**: Authors benefit from accurate assessment of their studies and are penalized for inaccurate assessment.
- **Organic funding mechanism**: Research funds (gain/loss from betting) move in the direction of smarter scientists and move away from bad/dishonest scientists.
- **Information alignment**: Information asymmetry between authors and community (P-hacking, multiple testing, etc.) priced into replication probabilities.
- **Reverse P-hacking**: Stated probability should be a conservative.

Possible objection:

- **Publishing everything will make science less reliable.**

  No, it will make it more reliable.

  - *Authors can't filter out their own bad findings.*
  - *Editors/reviewers can't suppress/censor "bad" ideas. Community must demonstrate that a claim is false by betting the other side and holding up during replication.*

H. Crane and R. Martin. (2018). In peer review we (don't) trust: How peer review's filtering poses a systemic risk to science. https://researchers.one/article/2018-09-17

Comparison: Prediction Markets vs. FPP

	Prediction Market	FPP
Replication Probability	Market-driven	Set by authors
Role of Probability	Publication Decision, Signaling	Price of real bets Skin in the game
Consequence: Prob too high	Type I error: Publication Author benefits	Author loses money/funds Challenger gains
Consequence: Prob too low	Type II error: Rejection Author penalized	Conservative estimate No harm, no foul
Author incentive	Overstate probability Deception	Understate probability Conservative
Information asymmetry	Between Market & Author	Price in or risk ruin
Challenges to paradigm	Upper hand for “normal science” New ideas suppressed	Equal footing with *status quo*

Final Remarks

I'm talking about a specific, extra type of integrity that is [beyond] not lying, but bending over backwards to show how you're maybe wrong, that you ought to have when acting as a scientist. (Feynman)

- Foundations of Probability (through FPP) can make science more reliable without assuming trust or integrity.
- Scientists who overstate importance/reliability of results go "out of business".
- Remove all barriers to publication: Publication no longer prestigious and therefore no longer the "target" of scientific research. (Goodhart's Law)

H. Crane. (2018). The Fundamental Principle of Probability. https://www.researchers.one/article/2018-08-16

H. Crane and R. Martin. (2018). In peer review we (don't) trust: How peer review's filtering poses a systemic risk to science. https://researchers.one/article/2018-09-17