Papers have bugs — what is to be done?

Enka Blanchard*[1,2] and Lê Thành Dũng (Tito) Nguyễn†[3]

[1] Laboratory of industrial and human automation, mechanics and computer science, Polytechnic University Hauts-de-France
[2] CNRS Center for Internet and Society, UPR 2000
[3] LIP, École Normale Supérieure de Lyon, France

1 Introduction: wrong science is unavoidable

In most sciences, it is universally acknowledged that something being published does not make it true. As shown by the replication crisis, this is not philosophical nitpicking but comes from errors with lasting impacts. Fields such as mathematics or theoretical computer science, on the contrary, seem to offer in principle the promise of eternal certainty: a proof of a theorem stays valid forever. And yet, every experienced researcher has run into wrong proofs – or even wrong statements[1] – in the published literature.

Faced with this state of affairs, one might be tempted to either blame the social organisation of research and call for reforming our practices, or seek to implement stricter quality controls before publication. While both may be desirable, we argue that they are not sufficient: we also need *self-correction mechanisms* that allows wrong science to be amended *a posteriori*.

An ecosystem promoting bad science. Awareness of the perverse incentives of academia is widespread by this point. "Publish or perish" implies a focus on maximising the number of papers accepted into prestigious venues. This means cutting corners on the science itself, the reviewing (of other's papers) as well as repeatedly submitting one paper to increases its chances to get published (as acceptance is partially random [CL21]), compounding some issues in the review process. In some fields (especially biomedical sciences), frequently used metrics such as the impact factors can even have a negative correlation with the quality of the research published [DH22, BB23]. Although (theoretical) computer scientists seem to give less credence to such metrics, we still follow a search for bureaucratically legible "excellence", such as the prestige hierarchy of conferences — through CORE rankings for example — in which the proportion of rejected papers is a relevant metric, independently of the quality of accepted (or rejected) papers. Even the simple task of avoiding predatory journals (and the "science" published therein) is becoming complex due to journal hijacking and takeovers [Boh15].

This brings us to another well-known structural issue, specific to computer science: our publishing system favours conference proceedings over journals. This pushes authors to rush to complete papers on a deadline[2] and PC members to rush for their own deadline, without taking the time to perform in-depth reviews. The end result is hastily written proofs in appendices that nobody reads.

A priori quality control does not suffice. As the quality of published proofs goes down, we must remember that it is only a matter of degree, as errors have always been published in reputable venues by respected researchers, even in the slower fields of mathematics (see e.g. [Buz20]). This seems to result inevitably from the increasing complexity and scale of the research done.

*The authors contributed equally and are listed in alphabetical order.

†Supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

[1] For instance, the second author discovered, with some colleagues, that the main theorem of a paper published in 2020 in the prestigious conference *Logic in Computer Science* (LICS) was in fact completely false.

[2] This was already denounced by Koblitz in 2007 [Kob07]. For a lighter take on it, see `https://youtu.be/707tBQ07I-A`.

This has prompted interest in verification of formal proofs by computers from mathematicians such as Kevin Buzzard or Vladimir Voevodsky, sometimes spurred by their own experience of publishing wrong results [Rod21, BL21]. Formalising a complex proof can significantly increase confidence in its correctness, by reducing it to the correctness of a small[3] proof-checking program. But while we are lucky to have such powerful tools at our disposal — many academic fields do not — they do not guarantee that the statement that has been proved (formally specified on the computer) is what we want (based on our informal understanding)[4].

The issue of interpretation remains, whether the proofs are written in a formal way or in mathematical vernacular. This raises concerns for proofs of theorems with practical applications, such as in cryptography, where modelling and complexity-theoretic assumptions can make some proofs irrelevant, as critiqued by Koblitz and Menezes [KM19].

2 Self-correction mechanisms

Current practices. A common idea in our research communities is that science is self-correcting, wrong results being progressively eliminated. Sadly, this is only partially true. Even when papers are known to be false, with published retractions, they continue being cited. This is known in other fields [TdSD17] but there is no reason to believe that we are impervious to such issues, as we also struggle to discourage the continued use of datasets featuring wrong data or ethical issues (such as lack of consent from experimental subjects), especially if they become benchmarks [KA22].

How do we handle wrong results? If the claims are important enough, they can lead to (failed) attempts at replications, critiques or counterexamples, although the process can take years [LC97]. A refutation can also be published without retracting the original article (or adding a warning on it) [Str19]. Sometimes, communities maintain folklore knowledge of which papers to avoid, or which proofs to be careful with. However, this esoteric knowledge has exclusionary effects: increasing the costs of entry and making collaborations harder, especially for junior/minority researchers. Furthermore, despite multiple initiatives that go in the right direction (e.g., OpenReview, PubPeer, Peer Community In), we do not have an established bug reporting mechanism (besides messaging authors which depends on their goodwill — and even then, they have no way to officially retract conference papers). The reliance on preprint servers (especially within the context of the fight for open access) worsens this issue, with 116 currently available proofs of P vs NP, including at least 6 published in peer-reviewed venues[5].

A journal of critique and refutations? To help address this ongoing issue, we propose the creation of a journal focused on refutation. Although some refutations do get published in generalist journals, they mostly target important results, leaving the possibility of accumulating many minor errors in certain fields. The journal's underlying idea is to provide a space for researchers to focus on bugfixing, cleanup and maintenance, while still "publishing" (and thus create professional incentives to do this work without requiring a full overthrow of current incentive structures). It would bridge the gap between the rare published refutations and the (often rigorous) arguments on community websites (PubPeer or StackExchange) which do not share the peer-reviewed status of the original articles and are sometimes discarded for this reason. We have explored this option since 2022 with colleagues from many fields and have ideas but also ongoing debates. First, the question of disciplinarity, as the journal would probably not have enough submissions to be focused on one field initially (one option is to create an anti-disciplinary journal that integrates the possibility of eventually splitting in specialised subjournals). Linked to this issue is the threshold criteria for refutations: should we only allow refutations of the most egregious errors, those which require limited expertise to be verified? Otherwise, how could we find the required expertise to analyse complex rebuttals? Moreover, should we allow for anonymous authors, and how would we handle accountability in this case — besides having open reviews and making the reviewers and editorial board accountable?

We submit this project to the community for feedback before going any further.

[3] Coq and Lean have huge code bases but the only sensitive part that needs to be trusted is a small "kernel".

[4] The Lean community is aware of this; see e.g. `https://leanprover-community.github.io/blog/posts/lte-examples/`

[5] As indexed on `https://www.win.tue.nl/~wscor/woeginger/P-versus-NP.htm`.

References

[BB23] Enka Blanchard and Zacharie Boubli. Recherche et dogmatisme : de l'improductivité du productivisme. *Questions de Communications*, 2023. URL: https://journals.openedition.org/questionsdecommunication/29994.

[BL21] Enka Blanchard and Giuseppe Longo. From axiomatic systems to the dogmatic gene and beyond. *Biosystems*, 204, 2021. doi:10.1016/j.biosystems.2021.104396.

[Boh15] John Bohannon. Feature: How to hijack a journal. *Science*, 2015. doi:10.1126/science.aad7463.

[Buz20] Kevin Buzzard. The future of mathematics?, 2020. Slides of a talk given at the Formal Methods in Mathematics workshop. URL: https://www.andrew.cmu.edu/user/avigad/meetings/fomm2020/slides/fomm_buzzard.pdf.

[CL21] Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: revisiting the 2014 NeurIPS experiment, 2021. arXiv:2109.09774.

[DH22] Michael R Dougherty and Zachary Horne. Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences. *Royal Society Open Science*, 9(8), 2022.

[KA22] Os Keyes and Jeanie Austin. Feeling fixes: Mess and emotion in algorithmic audits. *Big Data & Society*, 9(2), 2022.

[KM19] Neal Koblitz and Alfred Menezes. Critical perspectives on provable security: Fifteen years of "another look" papers. *Advances in Mathematics of Communications*, 13(4):517–558, 2019. doi:10.3934/amc.2019034.

[Kob07] Neal Koblitz. The uneasy relationship between mathematics and cryptography. *Notices of the AMS*, 54(8):972–979, 2007.

[LC97] Hoi-Kwong Lo and Hoi Fung Chau. Is quantum bit commitment really possible? *Physical Review Letters*, 78(17):3410–3413, 1997.

[Rod21] Andrei Rodin. Voevodsky's unfinished project: Filling the gap between pure and applied mathematics. *Biosystems*, 204, 2021.

[Str19] Lutz Straßburger. On the decision problem for MELL. *Theoretical Computer Science*, 768:91–98, 2019. doi:10.1016/j.tcs.2019.02.022.

[TdSD17] Jaime A Teixeira da Silva and Judit Dobránszki. Highly cited retracted papers. *Scientometrics*, 110(3):1653–1661, 2017.