© 2024. Published by The Company of Biologists Ltd | Journal of Experimental Biology (2024) 227, jeb247383. doi:10.1242/jeb.247383

PERSPECTIVE

Advancing data honesty in experimental biology

Shahar Dubiner[1,*,‡] and Matan Arbel-Groissman[2,*]

ABSTRACT

The ease with which scientific data, particularly certain types of raw data in experimental biology, can be fabricated without trace begs urgent attention. This is thought to be a widespread problem across the academic world, where published results are the major currency, incentivizing publication of (usually positive) results at the cost of lax scientific rigor and even fraudulent data. Although solutions to improve data sharing and methodological transparency are increasingly being implemented, the inability to detect dishonesty within raw data remains an inherent flaw in the way in which we judge research. We therefore propose that one solution would be the development of a non-modifiable raw data format that could be published alongside scientific results; a format that would enable data authentication from the earliest stages of experimental data collection. A further extension of this tool could allow changes to the initial original version to be tracked, so every reviewer and reader could follow the logical footsteps of the author and detect unintentional errors or intentional manipulations of the data. Were such a tool to be developed, we would not advocate its use as a prerequisite for journal submission; rather, we envisage that authors would be given the option to provide such authentication. Only authors who did not manipulate or fabricate their data can provide the original data without risking discovery, so the mere choice to do so already increases their credibility (much like 'honest signaling' in animals). We strongly believe that such a tool would enhance data honesty and encourage more reliable science.

KEY WORDS: Best practice, Data integrity, Data manipulation, Fraud, Reproducibility, Scientific misconduct

Introduction

Raw data from biological experiments can often be fabricated without trace, meaning that false results can be published with alarming ease. Although the publication of scientific papers is clearly beneficial for the scientific community and humanity as a whole, the way in which the publication process operates has created unique problems (Hausmann and Murphy, 2016). Published papers are almost the only currency in the academic world, and negative results are hardly ever published (Blanco-Perez and Brodeur, 2020; Fanelli, 2012), resulting in an abundance of fraudulent results circulating in the literature (Carafoli, 2015; Fanelli, 2009). A certain percentage of the scientific community, when exposed to temptation of the right incentive, will always be tempted to 'cut corners' and apply less than rigorous scientific methods or, in extreme cases, will even fabricate their results (Fang et al., 2012; Li et al., 2021). As an example, authors of influential studies in physiology and ecology were reported to have 'committed fabrication and falsification' of data, and a number of other papers are now facing scrutiny (see Enserink, 2021, 2022; López Lloreda, 2023). We can attest from personal experience that raw data that were undoubtedly false (including duplicated and impossible values) were found in papers sent to us for peer review; whether these were naïve errors or deliberate manipulations, we have no way of knowing. Thus, not only do false data exist in our discipline, but the effort needed to generate and even publish them is small compared with the great effort needed for their detection.

The ease and inconsequence with which data can be manipulated could conceivably tempt even a scientist with no prior intention to do so or cause an honest error to go unnoticed and find its way into publication. Even if cases of outright fraud are outliers in the scientific literature, their prevention should be improved, and in the case of genuine mistakes, the ease of data manipulation is enough to result in the unintentional circulation of unreliable data. Published papers are the milestones by which we measure researchers. Given an inability to detect fraudulent data, an unscrupulous 'cheater' will, like any cheater in a biological scenario, thrive at the expense of their honest peers, exacerbating the problem (Wade and Breden, 1980). This is an inherent blind spot in the way we conduct science and how we judge scientists. The problem of 'publish or perish' (Roland, 2007) has a secret third option: 'or simply cheat'. Thus, combating and eradicating fraud in our field of science (and science in general) requires fundamental changes to the norms and practices we take upon ourselves.

In this Perspective, we present our idea for a data tool that would promote data honesty and could push research to be more ethical and responsible. Many methods and devices currently used in experimental biology (e.g. microscopes, gel blots, imaging devices, microplate readers, spectrophotometers) provide output files that are completely amenable to changes or manipulation, and can also be fabricated in many cases (e.g. any spreadsheet file). Our aim is to encourage the development of a non-modifiable, authenticated data format that tracks deviations from the original output. These files could be added to the relevant supplement or repository when publishing papers, as a mark of honesty and fairness, thus enabling reviewers and readers to follow the logical footsteps of the authors and detect flaws in their handling of the data. We believe that such a tool, alongside the new and increasingly common methods for incentivizing honest results (Nosek and Lakens, 2014; Raff, 2013), would enhance data honesty and encourage more reliable science.

Authors should have the opportunity to prove their data is authentic

Current attempts to combat fraudulent data, and proposed methods of doing so, focus on the detection of inappropriately manipulated data during the review process (Boetto et al., 2020; Farid, 2006) and in already published works (Fanelli, 2009; Li et al., 2021). However, the prevention of data manipulation is still based on faith in the good practices and intentions of the scientists involved

[1]School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel. [2]The Shmunis School of Biomedicine and Cancer Research, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel.
*These authors contributed equally to this work

‡Author for correspondence (dubiner@mail.tau.ac.il)

S.D., 0000-0003-1771-0087

(Fanelli, 2009). Many journals now require detailed methodology for publication and require the raw data to be made available (or encourage it in certain cases, e.g. editorials by Caetano and Aisenberg, 2014; Franklin and Hoppeler, 2021; Ihle et al., 2017). Yet, this system is not resistant to cheating, as noted above. The self-policing nature of the scientific publication process is counterbalanced by the incentive to produce results, sometimes at the cost of (intentionally or inadvertently) weakening the scientific rigor of the research (Roland, 2007; Woolf, 1986). It is inefficient, and often unrealistic, to leave the entire task of data inspection in the hands of reviewers and editors (Triggle and Triggle, 2007), who even in the best scenarios can only test whether the results are supported by the data, not the data themselves. The welcome efforts to increase transparency in the presentation of methodology (Mebane et al., 2019; Sumpter et al., 2023) and data (Caetano and Aisenberg, 2014; Ihle et al., 2017; Roche et al., 2022) have yet to be extended to include transparency in whether and how the raw data were manipulated to create the published version, a possibility that we discuss in this Perspective. Note that we are not advocating for data policing. We do not think authentication should be a prerequisite for submission to journals, especially as this may deter submissions and disproportionally affect researchers from low-resource facilities. Rather, we advocate that authors should be given the opportunity to provide such authentication if they so wish.

Data honesty can also reduce human error

Human biases and errors are as much a cause of unreproducible scientific papers as outright cheating (Brown et al., 2018; May, 2021; Mebane et al., 2019). Eradicating error from research to the best of our ability is a major hurdle in any scientific endeavor, but its detection is often difficult. As long as the data themselves cannot be verified, the ability of a reader or reviewer (or even a collaborator on the same project) to verify that there was no bias or error in the data analysis process is almost non-existent. Even though the publishing of open data in experimental biology has surged in recent years, many of these data sets are incomplete or uninterpretable (approximately 50%; Roche et al., 2022), and none are fully verifiable by an external reader. By supplying the kind of authenticated raw data that we are advocating for in this Perspective – and hopefully including a step-by-step explanation of the data manipulation performed during the analysis process – a responsible reviewer or careful PI could double check that the analysis was devoid of bias or error. Software that would allow simple documentation of data and its analysis would also benefit the researcher by facilitating error detection and allowing transparency, thus limiting human mistakes in the scientific process.

Proposing a straightforward solution

Methods for protecting data integrity have been proposed in the past (Boetto et al., 2020; Li et al., 2021) but, perhaps because their relative intricacy makes them unsuitable for rapid and intuitive application, such solutions are not yet widely integrated into scientific norms. A variety of these previous solutions focus on already published data, and the rest – such as registered reports (Nosek and Lakens, 2014) – focus on the methodology prior to data collection. However, we believe that the intermediate stage (during data collection) is where authentication would be the most effective. Conceivably, all three approaches could work sequentially, in synergy, ensuring honesty and transparency throughout the scientific process.

As we see it, the most straightforward way to tackle the task of data authentication would be through implementing a universally trusted 'original version' format for the raw output from laboratory equipment. Many devices commonly used in experimental biology provide output as editable image files (e.g. microscopes, gel blots, imaging devices) or spreadsheets (e.g. metabolic systems, temperature loggers, microplate readers, spectrophotometers, fluorescence-activated cell sorting). At present, the authentication of images is labor intensive, not universally practiced, and 'vulnerable to a host of counter-measures' (Farid, 2006), whereas detecting forgery in spreadsheets such as CSV files can be extremely difficult, albeit not impossible (Simonsohn, 2013). However, if an incorruptible version of these files were to be generated alongside the normal output, in a reliable and secure way, this could demonstrate the integrity of the data provided by the authors, with minimal need for active investigation by the editors, reviewers or readers.

A basic and rapidly applicable approach would be a digitally 'watermarked' and non-editable read-only original version of the raw data. This could either be a file submitted as a supplement (or uploaded to the same repository as the edited data), or a copy immediately created online, which could be linked to the data during the publication process, but not edited. Any legitimate change or manipulation procedure could be openly explained in the text; any unexplained change will stand out as a discrepancy. Although our knowledge of software development is limited, we assume that such software would not need to interact with all existing laboratory equipment, but only with the relevant output (at the moment of output), which tends to be limited to a handful of file formats.

A more complicated, but more beneficial, variation would include automatically tracked changes (possibly including the date and time of each change), from the moment of data creation until their final submission. This would enable the reviewers – and readers – to carefully follow the changes to the file and allay any suspicion of inappropriate data manipulation should it arise. This could be developed as a separate program that could be linked to the relevant existing software, thus providing confirmation of authenticity. We call for the development of such software, and we believe that the effort required for its creation is low relative to its immense potential to benefit biological research worldwide.

With great responsibility

The scientific community has long entrusted the validity of scientific research to the peer-review process, meaning that researchers, on top of their routine work, are responsible for safeguarding the literature from scientific misdemeanor. Thus, unless the peer review system is drastically altered, to promote reproducibility in experimental biology, we should provide reviewers with the best tools possible to allow them to focus their limited time on the essentials of the manuscript and not on its technicalities. If we aim to enhance integrity and reproducibility in experimental biology, then providing tools for reviewers to do so, like those suggested in this Perspective, is of utmost importance.

The mental load currently carried by every research student and PI involves juggling experimental work with writing, reading and teaching; it is therefore naïve to believe that all researchers tasked with reviewing papers will (without being properly paid for their time) put in the necessary effort to do so thoroughly (Triggle and Triggle, 2007). Reviewing a paper involves assessing the scientific merits of the work and finding errors in the experimental work, its analysis, the writing of the paper and the conclusions reached from the data collected and analyzed. It is unreasonable to expect researchers, especially early career ones, still working to achieve their standing in the community, to shoulder all the grunt work of reviewing the underlying raw data as well. It might be argued that,

just as reviewers often neglect to examine the underlying data (Berberi and Roche, 2022, 2023) or pre-registration (Syed, 2024 preprint), they may fail to examine mismatches between the data and our proposed authenticated version. However, we see two ways in which our proposal overcomes this argument. In cases where fraud is already suspected and is being investigated, an authenticated data file would be extremely useful in providing the evidence so lacking in the majority of these investigations. Moreover, because providing authenticated data presents no risk to honest scientists, the very choice to do so could serve as a form of 'honest signaling' (a term we borrow from behavioural ecology; Grafen, 1990) of the data quality to the reviewer.

Conclusion

Despite our trust in the overall integrity of scientists, we are aware that research is nonetheless susceptible to fraud (Fanelli, 2009; Li et al., 2021), and we cannot assume our own field of research to be any exception. We argue that this problem could be ameliorated by enabling authors to easily authenticate their own raw data, as a form of 'honest signaling', to lend it immediate credibility. In order to prevent its corruption, such a signal needs to be impossible for a cheater to fake, in scientific publishing as much as in conventional biological signaling (Dawkins and Guilford, 1991; Wade and Breden, 1980). Unfortunately, our current non-verifiable data files are open to cheating. The data tool suggested in this Perspective would be especially well-suited in the context of replication, particularly for replications on highly controversial or debated topics (e.g. Clark et al., 2020; Clements et al., 2022, and the associated replies). If data authentication is presented by a journal as an option to authors, then we predict that, even without it being mandatory, this practice will gradually become the norm. In this scenario, data integrity will improve, reinstating trust in scientific experiments and the scientists who conduct them.

Competing interests

The authors declare no competing or financial interests.

Funding

S.D. is supported by the Azrieli Graduate Studies Fellowship.

References

Berberi, I. and Roche, D. G. (2022). No evidence that mandatory open data policies increase error correction. *Nat. Ecol. Evol.* **6**, 1630-1633. doi:10.1038/s41559-022-01879-9

Berberi, I. and Roche, D. G. (2023). Reply to: recognizing and marshalling the pre-publication error correction potential of open data for more reproducible science. *Nat. Ecol. Evol.* **7**, 1595-1596. doi:10.1038/s41559-023-02142-5

Blanco-Perez, C. and Brodeur, A. (2020). Publication Bias and Editorial Statement on Negative Findings. *Econ. J.* **130**, 1226-1247. doi:10.1093/ej/ueaa011

Boetto, E., Golinelli, D., Carullo, G. and Fantini, M. P. (2020). Frauds in scientific research and how to possibly overcome them. *J. Med. Ethics* **47**, e19. doi:10.1136/medethics-2020-106639

Brown, A. W., Kaiser, K. A. and Allison, D. B. (2018). Issues with data and analyses: errors, underlying themes, and potential solutions. *Proc. Natl. Acad. Sci. USA* **115**, 2563-2570. doi:10.1073/pnas.1708279115

Caetano, D. S. and Aisenberg, A. (2014). Forgotten treasures: the fate of data in animal behaviour studies. *Anim. Behav.* **98**, 1-5. doi:10.1016/j.anbehav.2014.09.025

Carafoli, E. (2015). Scientific misconduct: the dark side of science. *Rendiconti Lincei* **26**, 369-382. doi:10.1007/s12210-015-0415-4

Clark, T. D., Raby, G. D., Roche, D. G., Binning, S. A., Speers-Roesch, B., Jutfelt, F. and Sundin, J. (2020). Ocean acidification does not impair the behaviour of coral reef fishes. *Nature* **577**, 370-375. doi:10.1038/s41586-019-1903-y

Clements, J. C., Sundin, J., Clark, T. D. and Jutfelt, F. (2022). Meta-analysis reveals an extreme "decline effect" in the impacts of ocean acidification on fish behavior. *PLoS Biol.* **20**, e3001511. doi:10.1371/journal.pbio.3001511

Dawkins, M. S. and Guilford, T. (1991). The corruption of honest signalling. *Anim. Behav.* **41**, 865-873. doi:10.1016/S0003-3472(05)80353-7

Enserink, M. (2021). Sea of doubts. *Science* **372**, 560-565. doi:10.1126/science.372.6542.560

Enserink, M. (2022). Star marine ecologist guilty of misconduct, university says. *Science* **377**, 699-700. doi:10.1126/science.ade3374

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* **4**, e5738. doi:10.1371/journal.pone.0005738

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891-904. doi:10.1007/s11192-011-0494-7

Fang, F. C., Steen, R. G. and Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proc. Natl. Acad. Sci. USA* **109**, 17028-17033. doi:10.1073/pnas.1212247109

Farid, H. (2006). Exposing digital forgeries in scientific images. In: Proceedings of the 8th workshop on Multimedia and security. doi:10.1145/1161366.1161374

Franklin, C. E. and Hoppeler, H. H. (2021). Elucidating mechanism is important in forecasting the impact of a changing world on species survival. *J. Exp. Biol.* **224**, jeb242284. doi:10.1242/jeb.242284

Grafen, A. (1990). Biological signals as handicaps. *J. Theor. Biol.* **144**, 517-546. doi:10.1016/S0022-5193(05)80088-8

Hausmann, L. and Murphy, S. P. (2016). The challenges for scientific publishing, 60 years on. *J. Neurochem.* **139**, 280-287. doi:10.1111/jnc.13550

Ihle, M., Winney, I. S., Krystalli, A. and Croucher, M. (2017). Striving for transparent and credible research: practical guidelines for behavioral ecologists. *Behav. Ecol.* **28**, 348-354. doi:10.1093/beheco/arx003

Li, W., Bordewijk, E. M. and Mol, B. W. (2021). Assessing research misconduct in randomized controlled trials. *Obstet. Gynecol.* **138**, 338-347. doi:10.1097/AOG.0000000000004513

López Lloreda, C. (2023). University investigation found prominent spider biologist fabricated, falsified data. *Science*. doi:10.1126/science.adi6906

May, J. (2021). Bias in science: natural and social. *Synthese* **199**, 3345-3366. doi:10.1007/s11229-020-02937-0

Mebane, C. A., Sumpter, J. P., Fairbrother, A., Augspurger, T. P., Canfield, T. J., Goodfellow, W. L., Guiney, P. D., LeHuray, A., Maltby, L., Mayfield, D. B. et al. (2019). Scientific integrity issues in environmental toxicology and chemistry: improving research reproducibility, credibility, and transparency. *Integr. Environ. Assess. Manag.* **15**, 320-344. doi:10.1002/ieam.4119

Nosek, B. and Lakens, A. D. (2014). Registered reports. *Soc. Psychol.* **45**, 137-141. doi:10.1027/1864-9335/a000192

Raff, J. W. (2013). The San Francisco declaration on research assessment. *Biol. Open* **2**, 533-534. doi:10.1242/bio.20135330

Roche, D. G., Raby, G. D., Norin, T., Ern, R., Scheuffele, H., Skeeles, M., Morgan, R., Andreassen, A. H., Clements, J. C., Louissaint, S. et al. (2022). Paths towards greater consensus building in experimental biology. *J. Exp. Biol.* **225**, jeb243559. doi:10.1242/jeb.243559

Roland, M. (2007). Publish and perish. *EMBO Rep.* **8**, 424-428. doi:10.1038/sj.embor.7400964

Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychol. Sci.* **24**, 1875-1888. doi:10.1177/0956797613480366

Sumpter, J. P., Runnalls, T. J., Johnson, A. C. and Barcelo, D. (2023). A 'Limitations' section should be mandatory in all scientific papers. *Sci. Total Environ.* **857**, 159395. doi:10.1016/j.scitotenv.2022.159395

Syed, M. (2024). Some data indicating that editors and reviewers do not check preregistrations during the review process. *PsyArXiv Preprints*. doi:10.31234/osf.io/nh7qw

Triggle, C. R. and Triggle, D. J. (2007). What is the future of peer review? Why is there fraud in science? Is plagiarism out of control? Why do scientists do bad things? Is it all a case of: "All that is necessary for the triumph of evil is that good men do nothing?". *Vasc. Health Risk Manag* **3**, 39-53.

Wade, M. J. and Breden, F. (1980). The evolution of cheating and selfish behavior. *Behav. Ecol. Sociobiol.* **7**, 167-172. doi:10.1007/BF00299360

Woolf, P. K. (1986). Pressure to publish and fraud in science. *Ann. Intern. Med.* **104**, 254. doi:10.7326/0003-4819-104-2-254