Written Evidence Submitted by the University of Oxford

(RRE0076)

Aksel Braanen Sterri,[1,2] Rebecca Brown,[1] Brian Earp[1,3], Julian Savulescu[1]

1. Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Suite 8, Littlegate House, St Ebbes Street, Oxford OX1 1PT, United Kingdom
2. Faculty of Health Sciences, Oslo Metropolitan University, 0166 Oslo, Norway
3. Yale University Departments of Psychology and Philosophy, New Haven, CT 06511 USA

We are academics from the University of Oxford who work on the ethics of science and medicine. We are concerned with ensuring that scientific research is conducted ethically. This includes ensuring that it effectively contributes to the public good of scientific knowledge. In this submission, we address most topics in the call for evidence.

There is evidence that, due in large part to structural issues in the science-to-publication pipeline, many impactful studies published in highly respected journals fail to replicate. This means that, when those studies are re-run, they do not find the same results. Since we do not know which (other) studies would hold up under closer scrutiny, we are not in a position to know how much trust in the scientific literature in many areas is warranted. Since many important decisions in politics, business, medicine, and people's private lives are based on scientific studies, this so-called "replication crisis" has enormous societal costs. One study puts the annual costs at $28 billion (Freedman, Cockburn, Simcoe 2015). There are at least four main sources of concern.

1. On average, depending on the discipline or sub-discipline, perhaps as much as 50 percent of published research findings fail to replicate when independent labs attempt to repeat the underlying experiments (Monya 2015, Fidler and Wilcox 2021).
2. Traditional measures for assuming that a published paper meets a baseline level of trustworthiness, such as the impact factor of the journal in which it is published and its citation count, are not reliable indicators of trustworthiness (Brembs 2018).

3. Replication is attempted for proportionally few papers, even fewer papers are retracted, and papers continue to be cited despite being retracted or failing to replicate (Redman et al. 1998, Yang et al. 2020). Thus, we do not know, by and large, which individual studies are to be trusted.

4. There is a large amount of evidence for publication bias where pharmaceutical companies and other stakeholders fail to publish negative results which may frustrate their interests (Savulescu and Chalmers 1995)

Despite these problems, it may not be the case that researchers are entirely unable to determine the trustworthiness of certain studies, in terms of their replicability. When researchers are asked to place bets on studies according to the likelihood of replication (in so-called prediction markets), they are very accurate. Prediction markets involve drawing upon the 'wisdom of the crowds' and they reveal that researchers, collectively, know which studies to trust and not (Dreber et al. 2015, Camerer et al. 2018, Gordon et al. 2020, de Menard 2020).

Why, then, are so many untrustworthy studies still published? A major problem is the incentive structure in the science-production complex: In short, the reason why scientists cut corners (or rely on suboptimal research design or statistical methods) is because they benefit from doing so and the reason corrective measures have not been adequate for stemming the tide of poor research is that scientists who spend their time attempting to address such issues do not professionally benefit from doing so. Funders, universities, journals, researchers, all contribute to uphold the current incentive structure. More specifically:

1. Careful and trustworthy science is difficult and thus takes more time and resources.

2. Careers, grants, and status depend on frequent publications in highly ranked journals and only to a lesser extent on following best practice.

3. Novelty is rewarded, and replications are not considered novel. Researchers doing replications might thus be considered less talented and even as potential troublemakers.

4. Journal editors and publishing companies want their journal(s) to be considered credible (which provides an incentive to scrutinize research) but they also want attention (which provides an incentive not to).

5. Peer reviewers, the supposed main guardians of good research, have little time and power. They are not rewarded for their time and therefore spend less time on other people's papers than would perhaps be ideal. Moreover, if peer reviewers report questionable research practices, this will not become public – the researchers in question can just try another journal. (Researchers' reputation might, however, be damaged with the editors of the journal.)

6. Funders of research, particularly pharmaceutical companies, have conflicts of interest and bias against negative results.

Aims

When thinking about how to improve upon the current system, there are several concerns that need to be kept in mind. We want researchers to:

1. spend time on important research,
2. produce high quality research which is published regardless of whether the findings are positive or negative,
3. publish findings that hold up under closer scrutiny
4. correct flawed research before it has an impact
5. spread credible and important research across the research community and to the wider public
6. be certified and rewarded if they produce credible and important research (with positions, grants, airtime, impact on policy)

What is likely to help?

1. *Good research practice requirements for publications and when applying for grants*: Pre-registration of studies, making publication of results publicly available on completion, sharing data and code, etc., will make p-hacking, publication bias, and other problematic practices more difficult.

2. *Improved pre-publication testing at journals:* Currently, few resources are spent on independent testing of the research that gets published. If journals spent more resources on this, more flaws and dubious research practices would be corrected pre-publication.

3. *Replication of published research where appropriate*: Makes it more likely that poor research practices will be detected, making it less rewarding to cut corners.

Replication could also provide important information to the research community and the wider public about which studies to trust and not.

4. *Updating public records*: For replications to be a deterrent to suboptimal research practices and provide requisite information, the results of replications must be spread widely (Redman et al. 1998). One suggestion is that articles that fail to replicate could have that clearly marked in the online version of the original journal article: "this article has failed to replicate" (with citations and links to the non-replications and any replies by the authors).

5. *Reward honest researchers:* Universities, funders, researchers, policy makers, and journalists could take failed or successful replications into account as one, among several, factors.

6. *Reward high-quality research, not quantity or striking positive results:* When assessing candidates for grants and career advancement, one idea is to assess fewer research contributions. To avoid journals publishing research that is "too good" to be true, one could impose negative sanctions on journals that publish research that often fails to replicate.

Potential downsides

1. Pre-registration might just change p-hacking and other research practices to the pre-pre-registration stage. Sometimes sharing of data is difficult for privacy concerns.

2. Replication is not the only thing we care about. A result can hold up but nevertheless be trivial. We want important and ground-breaking research to be pursued and these come with a high risk of failure. There is a risk that more replications will lead to less risky science, rather than the same science being pursued in a more thorough way.

3. "Naming and shaming" can exacerbate this tendency, by creating a chilling effect on new, experimental research. It is difficult to break new ground and we should therefore expect the most ground-breaking results to be wrong.

4. Replications and "naming and shaming" can have a harmful effect on honest researchers if we conflate failure to replicate with poor quality or deceptive research practices. It can therefore be more effective and less harmful to shame journals and publishing houses that publish poor research, rather than individual researchers.

5. If one retracts studies that fail to replicate or stop using studies to justify interventions, it can be harmful to replicate certain studies (where the harm of false positives is low, and the harm of false negatives is high). This could be the case for medicines with few negative side-effects and potentially huge benefits. Unnecessary replication of randomised controlled trials which have proven superiority of a particular treatment are harmful and should not be done (Savulescu and Chalmers 1995)

6. Replications have a substantial opportunity cost. If we replicate all studies, there are many good ideas that must be left on the table because of time and resource constraints (Everett and Earp 2015).

One Solution: Prediction markets

This shows that there are both benefits and downsides to replication. We therefore need a way to figure out which papers to replicate and which to leave untouched. Above we mentioned the power of prediction markets in revealing the collective knowledge in the research community. One idea is to create such a prediction market on a bigger scale, and pool (relevant, appropriately credentialed) researchers' predictions on which papers they believe would hold up in a replication. One could use these predictions to identify research that should be prioritised for replication. The papers that get a low ranking but have a high citation count (and thus are considered important) would be good candidates for replication.

The prediction market would thus serve as a publicly available barometer of the credibility of individual studies (and through aggregation, individual fields) that can be used by journalists, politicians, NGOs etc. to navigate in the complex world of science. It would also provide information for as to particularly untrustworthy journals and/or publishing houses. It would thus provide an incentive to produce and publish credible research. The actual replications will, in turn, provide information to researchers and thus improve the accuracy of the prediction market (when researchers adjust their bets in the light of the new replication).

Literature

Baker, Monya (25 May 2016). 1,500 scientists lift the lid on reproducibility. Nature 533 (7604): 452–454. doi:10.1038/533452a.

Camerer, Colin F., Anne Dreber, Felix Holzmeister *et al.* (2018) Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour* 2**,** 637–644. https://doi.org/10.1038/s41562-018-0399-z

de Menard, Alvardo (2020). What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers. *Fantastic Anachronism* https://fantasticanachronism.com/2020/09/11/whats-wrong-with-social-science-and-how-to-fix-it/

Dreber, A, Thomas Pfeiffer, Johan Almenberg et al. (2015) Prediction markets in science. Proceedings of the National Academy of Sciences 112 (50): 15343-15347. doi:10.1073/pnas.1516179112.

Everett, Jim A. C. and Brian D. Earp (2015) A tragedy of the (academic) commons: interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in psychology*, *6*, 1152.

Fidler, Fiona and John Wilcox (2021) Reproducibility of Scientific Results. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.) https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/.

Freedman, Leonard P., Iain M. Cockburn, Timothy S. Simcoe (2015) The economics of reproducibility in preclinical research. PLoS Biology 13, e1002165.

Gordon, Michael, Domenico Viganola, Michael Bishop et al. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme *Royal Society Open Science* 7(7). https://doi.org/10.1098/rsos.200566

Redman, Barbara K., Hossein N. Yarandi, Jon F. Merz (2008) Empirical developments in retraction. *Journal of Medical Ethics* 34:807-809.

Savulescu, Julian, Iain Chalmers, Jennifer Blunt (1996) Are research ethics committees behaving unethically? Some suggestions for improving performance and accountability. *British Medical Journal* 313:1390–3.

Yang Yang, Wu Youyou, Brian Uzzi (2020), Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117 (20) 10762-10768; DOI: 10.1073/pnas.1909046117.

(September 2021)