Is there a digital File Drawer problem?

written by Jeremy Huggett | 16/04/2018

by Sailko via Wikimedia Commons CC BY-SA 3.0

Although there has been a dramatic growth in the development of autonomous vehicles and consequent competition between different companies and different methodologies, and despite the complexities of the task, the number of incidents remains remarkably small though no less tragic where the death of the occupants or other road users is involved. Of course, at present autonomous cars are not literally autonomous in the sense that a human agent is still required to be available to intervene, and accidents involving such vehicles are usually a consequence of the failure of the human component of the equation not reacting as they should. A recent fatal accident involving a Tesla Model X (e.g. Hruska 2018) has resulted in some push-back by Tesla who have sought to emphasise that the blame lies with the deceased driver rather than with the technology. One of the company’s key concerns in this instance appears to be the defence of the functionality of their Autopilot system, and in relation to this, a rather startling comment on the Tesla blog recently stood out:

> No one knows about the accidents that didn’t happen, only the ones that did. The consequences of the public not using Autopilot, because of an inaccurate belief that it is less safe, would be extremely severe. (Tesla 2018).

Discussions of the ethics surrounding autonomous vehicles and their decision-making priorities frequently refer to the ‘Trolley problem’ (is it ‘better’ to run over an elderly person or a young child, for instance?) but this Tesla statement suggests a different ethical problem. It implies that while they can cite the number of fatalities per million miles associated with their technology (1:320 compared to 1:86 across all vehicles) they are apparently not able to produce data relating to instances where the driver was required to intervene in order to prevent an incident. It's hard to believe they haven't collected these data at all, so presumably they are inaccessible in some way (for commercial reasons, perhaps) which means they may as well not exist as far as the wider world is concerned. This is a variant on the 'File Drawer problem' in which numerous studies may have been undertaken but are never reported, instead ending up in a file drawer gathering dust.

The 'File Drawer problem' is a term coined by Robert Rosenthal (1979) in a paper which expressed the view that, at its extreme, academic journals were filled with the 5% of studies that showed significant results, while the file drawers back in the lab were filled with the 95% of the studies that showed non-significant results, and consequently risked promulgating Type I errors (i.e. false positives). If true, this worst-case scenario where only significant studies are published and there are no published non-significant results would have considerable implications for the outcomes of research. Not only would the conclusions that could be drawn from such a set of studies be skewed, the implications for meta-analyses – such as are increasingly a feature of 'big data' studies – are considerable if the outcomes being factored in were primarily positive.

What would be the consequences for archaeology if we did not know about problems with our digital studies such that our conclusions were biased? I've commented on a number of occasions that utopianism is frequently a characteristic of archaeological discussions of digital tools (for example, Huggett 2004; 2012; 2015) which leads to a tendency to report positively. For instance, a typical digital workflow study (as is commonly found in CAA proceedings, for instance) almost by definition outlines a working procedure and in the process any negative aspects in terms of what had to be overcome tend to be downplayed (e.g. Huggett 2017, section 8.4). In general, we read about positive outcomes, and more rarely encounter negative results.

If, as seems more than likely, there is a lack of negative results in digital archaeology, then we are not alone in this. For example, Danielle Fanelli outlines what she describes as a realistic scenario, which includes:

> … various forms of conscious and unconscious biases that affect all stages of research—e.g., study design, data collection and analysis, interpretation and publication—producing positive findings when there should be none, thus creating distortions that are difficult to correct a posteriori (2012, 892).

Significantly for archaeology, she goes on to suggest that this situation will be particularly acute in fields where theories and methods are less clearly defined, and true replication is rare or impossible. Furthermore, she finds that the average frequency of positive results is higher in the social sciences and in applied versus pure disciplines (2012, 893; see also Fanelli 2010). Such results are not unconnected with the replication crisis in a number of these disciplines.

Fanelli suggests the high proportion of positive relative to negative results published could arise from one or more of three factors (2012, 899):

1. Researchers address hypotheses that are more likely to be confirmed in order to get

publishable results. This reflects a perception that publishers are biased toward publishing ‘positive’ results except where they provide strong evidence countering an established hypothesis. It seems unlikely that archaeological outlets are immune from this.

2. There is an increase in average statistical power of studies over time, perhaps though increasing the sample size (as could be the case with a big-data-style meta-analysis, for instance), although Fanelli finds no evidence to support such a claim. There are certainly growing numbers of meta-analyses being undertaken in the context of big data studies in archaeology.
3. Negative results could be less frequently submitted and/or accepted for publication, or turned into positive results “through post hoc re-interpretation, re-analysis, selection or various forms of manipulation/fabrication” (Fanelli 2012, 899) and although she finds no evidence that negative results were increasingly embedded in papers reporting positive outcomes, multi-hypothesis studies frequently began with a negative result at the outset before moving onto positive ones.

Of course, these issues aren’t restricted to digital archaeology – there are implications for the entire discipline. However, the digital is implicated in at least the sense that it is a facilitator: it spreads the risk through making increasing numbers of datasets available which may or may not be affected by confirmation bias, and subsequently supporting their incorporation within meta-analyses. However, the digital can also help address – or at least evaluate (and maybe alleviate) – the scale of the problem in a number of ways.

First, we should establish some idea of the extent of positive academic bias in our disciplinary structures. As Marco Pautusso outlines, this can be encountered along a spectrum, from selective funding though publication bias to citation bias (2010, 198), and it has potentially far-reaching implications beyond the obvious disciplinary questions. For instance, the demand for novel and innovative studies in the context of formal research assessments and reviews is a cause of significant stress and presents major challenges to personal well-being, especially if valid work is going unrecognised because, for whatever reason, it isn’t considered to be ‘significant enough’ to be worthy of publication. We should maybe consider the potential for journals of negative results, as have been published in other disciplines (e.g. Journal of Articles in Support of the Null Hypothesis, the Journal of Negative Results Ecology & Evolutionary Biology, the All Results series of journals in a number of disciplines, and the Journal of Pharmaceutical Negative Results).

Secondly, we should undertake a study of online publication databases in order to assess the extent to which the file drawer problem is an issue within archaeology. However, the methodologies outlined by Fanelli (2012) and Pautasso (2010), for example, may not necessarily map across to archaeology. For example, searching for *“test* the hypothes*”* in archaeology using Clarivate Analytics’ Web of Science citation index only generated 144 results while looking for *“no significant difference”* generated 52 hits, but in both cases a high proportion of the papers were not in archaeological journals as such. A more nuanced search strategy using the full published texts would likely be needed to extract data of archaeological value.

Thirdly, we should establish some means of evaluating whether publication biases exist, and, if so, whether they are significant or not. It is unclear (to me, at least!) whether methods for calculating the impact of the file drawer problem by estimating the potential for unpublished or missing studies to alter outcomes (e.g. Rosenthal 1979; Rosenberg 2005), can be easily applied within archaeology. However, with the increasing number of archaeological meta-analyses incorporating the results of multiple independent studies we surely need some means of evaluation in order to have confidence in their results.

Interestingly, there might be something to be said for **not** seeking to address these questions. Fanelli (2012, 900) notes that those disciplines showing the strongest increase in positive results were in those very disciplines which had attempted to correct the problem (for example, by registering clinical trials, enforcing standards of reporting, and creating journals of negative results). So it may indeed be that ignorance is bliss. However, even if this were to be the case, the significant questions associated with the individual personal and professional cost of such biases should warrant their investigation and resolution.

References

Fanelli, D. 2010 '"Positive" Results Increase Down the Hierarchy of the Sciences', *PLOS ONE* 5(4): e10068. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010068

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

Hruska, J. 2018 'Tesla Blames Driver in Model X Autopilot Crash', *ExtremeTech* (April 13, 2018) https://www.extremetech.com/extreme/267417-tesla-blames-driver-in-model-x-autopilot-crash

Huggett, J. 2004 'Archaeology and the new technological fetishism', *Archeologia e Calolatori* 15, 81-92. http://www.progettocaere.rm.cnr.it/databasegestione/open_oai_page.asp?id=oai:www.progettocaere.rm.cnr.it/databasegestione/A_C_oai_Archive.xml:360

Huggett, J. 2012 'What lies beneath: lifting the lid on archaeological computing', in: Chrysanthi, A., Murrietta Flores, P. and Papadopoulos, C. (eds.) *Thinking Beyond the Tool: Archaeological Computing and the Interpretative Process*. Oxford: Archaeopress, pp. 204-214. http://eprints.gla.ac.uk/61333/

Huggett, J. 2015 'A Manifesto for an Introspective Digital Archaeology', *Open Archaeology* 1, 86-95. https://dx.doi.org/10.1515/opar-2015-0002

Huggett, J. 2017 'The apparatus of digital archaeology', *Internet Archaeology* 44. https://dx.doi.org/10.11141/ia.44.7

Pautasso, M. 2010 'Worsening file-drawer problem in the abstracts of natural, medical and social science databases', *Scientometrics* 85, 193-202. https://dx.doi.org/10.1007/s11192-010-0233-5

Rosenberg, M. 2005 'The File-Drawer Problem Revisited: A General Weighted Method for Calculating Fail-Safe Numbers in Meta-Analysis', *Evolution* 59 (2), 464-468. http://www.jstor.org/stable/3448935

Rosenthal, R. 1979 'The "File Drawer Problem" and Tolerance for Null Results', *Psychological*

Bulletin 86 (3), 638-641. https://dx.doi.org/10.1037/0033-2909.86.3.638

Tesla 2018 ‘An Update on Last Week’s Accident’ Tesla blog (March 30, 2018) https://www.tesla.com/en_GB/blog/update-last-week%E2%80%99s-accident