Checklist for research ethics in Stat+ML

Aaditya Ramdas (aramdas@cmu.edu), Carnegie Mellon University

Clearly, many of the opinions below are controversial, and even evolving, opinions based on experence. But that doesn't mean that we cannot write something down, discuss it, and evolve how we think accordingly.

- **Scooping** is unfortunately somewhat uncommon, both intentionally (misconduct) and unintentionally (misfortune/misunderstanding). It may stem from memory/psychology, where you do not remember the source of your ideas and credit your own smarts for it. It may involve intention or incentives, often between people of differing power/influence. It may or may not ever come to light, and may or may not have repercussions.
- **Acknowledgments.** Acknowledging people in your paper/talk is never a bad thing. People tend to acknowledge famous people, and discount the suggestions of less famous people; this should be actively avoided. Anyone who has read a preprint (or practice talk) and provided reasonable feedback should be acknowledged. Early conversations that shaped the studied problem could also be acknowledged. You must definitely acknowledge (and cite) borrowed code that you extended.
- **Authorship.** To avoid misunderstandings, get authorship expectations sorted out early with clear in-person or written conversations. For papers that developed organically, here is a rough guide: for people who spent less than ten hours on the paper, an acknowledgment should suffice (but, do inform them). For people who may have spent 10-25 hours, you could offer authorship and they may choose to decline, unless they contributed central ideas. (These numbers assume that a first author spends $10^2 - 10^3$ hours on a paper.)
- **Author order.** IMO, alphabetical ordering is to be avoided if possible. It is well known to hurt some cultures more (Chinese last names often start with W, X, Y, Z). Papers are often referenced as A et al, making it easy to associate the work with A. It opens the door to uninentended psychological biases (like primacy effects) — studies show that awards and tenure rates are correlated to the first letter of the last name, and it is not a stretch to attribute some causality. My personal rule is simple: in a student+faculty (or postdoc+faculty) paper, the student is first author, since they have more to gain than me. In a paper without a student/postdoc, whoever "led" the writing (often formatting, submission, rebuttal) is first author. If multiple students contributed equally, use a * footnote to reflect this, and alternate order in followup papers.
- **Biased references.** (IMO) Famous authors, and authors from prestigious institutions, can sometimes unintentionally ignore or downplay the work of less famous authors or those from less-known institutions. This can happen in many ways: the usage of words like 'seminal' or 'classical', citing some authors by name (A and B) and others by number [23], or just completely ignoring their work and not citing them. These may happen unintentionally, or subconsciously, due to the slow permeation of subtle social cues. Attempt to avoid consciously and credit when it is due (handsomely if needed).
- **The appearance of unethical conduct.** Conflicts of interest (say, due to the agency funding the research) should be declared clearly. Parallel work should be cited as such. The appearance (to others) of unethical conduct is arguably as important as unethical conduct itself. If post-hoc evidence comes to light that you were aware of potential conflicts or of parallel work, and you chose not to disclose it, then there may be an increased burden of evidence to show that the negligence to not mention it did not arise out of a desire to hide (but out of forgetfulness, or supposed irrelevance, or priority of ideas).
- **Communicating expectations and worries.** The key to getting ahead of future potential issues is often pre-emptive transparent communication: with collaborators, with funders, with the public (in your paper), with the journal that you submit to, etc. Communicate clearly and trust them to respond fairly.
- **Internal signal.** Think about what would happen if all your emails and verbal conversations with your collaborators (pertaining to a particular paper) became public, perhaps provided by one of them. Would you be worried about how you will be judged? If yes, this is an internal signal that perhaps your flirting with (or crossing) the boundary of what is ethical and what is not. Step back and reassess, earlier rather than later.

Clearly, the above discussion is *far* from complete, and morally ambigous situations may arise in admissions, faculty hiring, paper reviewing and what not. Tune your internal signal to be reliable in these cases.

Since the above may not suffice as enough actionable advice, here is more:

1. If you go up to a person at their poster or after their talk, and chat with them broadly about future directions, that limited interaction usually does not involve any commitment from either side. But, if either person finds the brainstormed ideas interesting enough to pursue, it may be courteous to invite the other to participate (if you can track them down) or at the very least acknowledge them in the resulting work. This is not a symmetric interaction — it is more likely that the presenter is continuing research along that line of work, and so it is the attendee that is at a greater risk of "scooping" unintentionally.
2. If you spot a paper on arXiv and you like the idea, I would personally recommend not rapidly building on it immediately (within 2-3 months) even if you see some "easy" extensions. Short, quick followup papers are anyway rather unlikely to be high impact (they are derivatives, piggybacking off other peoples' good ideas). On the contrary, such a work is quite likely to annoy the original authors who may be taking their time to seek feedback and improve the paper on other angles before submission.
3. On the same note, I think that it is completely reasonable to protect yourself from getting scooped. Sometimes, I don't put up the entire codebase to reproduce all experiments until *after* a paper is accepted. Sometimes, I put a paper on arXiv couple of months before submission (to get feedback), sometimes just after submission (because why not), but sometimes I do it a few months after submission (if I think it's important work in an area that's a bit too hot and a young student I'm working with needs some time and protection from the previous point's vulnerability).

I am copying below some complementary recommendations from the American Statistical Association.

- **Professional Integrity and Accountability:** "The ethical statistician uses methodology and data that are relevant and appropriate; without favoritism or prejudice; and in a manner intended to produce valid, interpretable, and reproducible results."
- **Integrity of Data and Methods:** "The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may affect the integrity or reliability of the statistical analysis."
- **Responsibilities to Science/Public/Funder/Client:** "The ethical statistician supports valid inferences, transparency, and good science in general, keeping the interests of the public, funder, client, or customer in mind."
- **Responsibilities to Research Subjects:** "The ethical statistician protects and respects the rights and interests of human and animal subjects at all stages of their involvement in a project."
- **Responsibilities to Research Team Colleagues:** "Science and statistical practice are often conducted in teams made up of professionals with different professional standards. The statistician must know how to work ethically in this environment."
- **Responsibilities to Other Statisticians or Statistics Practitioners:** "The practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers drawing on their own unique sets of experiences can arrive at different and potentially diverging judgments about the plausibility of different explanations. Even in adversarial settings, discourse tends to be most successful when statisticians treat one another with mutual respect and focus on scientific principles, methodology, and the substance of data interpretations."
- **Responsibilities Regarding Allegations of Misconduct:** "The ethical statistician understands the differences between questionable statistical, scientific, or professional practices and practices that constitute misconduct."
- **Responsibilities of Employers, Including Organizations, Individuals, Attorneys, etc:** "Those employing any person to analyze data are implicitly relying on the profession's reputation for objectivity. However, this creates an obligation on the part of the employer to understand and respect statisticians' obligation of objectivity."