A Two-Sided Discussion of Preregistration of NLP Research

Anders Søgaard **Daniel Hershcovich**
Department of Computer Science
University of Copenhagen
{soegaard,dh}@di.ku.dk

Miryam de Lhoneux
Department of Computer Science
KU Leuven
miryam.delhoneux@kuleuven.be

Abstract

Van Miltenburg et al. (2021) suggest NLP research should adopt *preregistration* to prevent fishing expeditions and to promote publication of negative results. At face value, this is a very reasonable suggestion, seemingly solving many methodological problems with NLP research. We discuss pros and cons—some old, some new: a) Preregistration is challenged by the practice of retrieving hypotheses after the results are known; b) preregistration may bias NLP toward confirmatory research; c) preregistration must allow for reclassification of research as exploratory; d) preregistration may *increase* publication bias; e) preregistration may *increase* flag-planting; f) preregistration may *increase* p-hacking; and finally, g) preregistration may make us less risk tolerant. We cast our discussion as a dialogue, presenting both sides of the debate.

1 Preregistration

Should NLP researchers be required to preregister their studies? Van Miltenburg et al. (2021) present arguments *for* preregistration, recently echoed by Ulmer et al. (2022). Preregistration has its origin in preregistration of clinical trials,[1] and amounts to the following: Before you initiate a set of experiments, you register your hypotheses, your experimental design and how you plan to analyze your results. Registration is time-stamped on an online platform with general public access. You then follow your plan as closely as possible and report any divergences in your final publication.

The discussion in van Miltenburg et al. (2021) is not unprecedented. Preregistration has been debated in epidemiology (Lash and Vandenbroucke, 2012), social psychology (Veer and Giner-Sorolla, 2016), experimental economics (Strømland, 2019) and information systems research (Bogert et al., 2021). Our discussion is inspired by the discussion in epidemiology, which is similar to NLP in focusing on data analysis rather than clinical trials.

There is an important ambiguity in how preregistration is discussed: Is the preregistration entry peer-reviewed or not? Chambers (2019) sees preregistration as a peer-reviewed process, and this is also what van Miltenburg et al. (2021) suggest for NLP. We therefore assume peer-reviewed preregistration below. The required format of the registered report is also important. In their Appendix, van Miltenburg et al. (2021) provide example questionnaires. We will assume registered reports will be lists of answers to such questionnaires, but in §9, we will suggest a few revisions to the questions formulated by van Miltenburg et al. (2021).

2 Why Preregister NLP Research?

Van Miltenburg et al. (2021) present four reasons for adopting preregistration in NLP: distinguishing between confirmatory and exploratory research, avoiding fishing expeditions and harking, mitigating publication bias and avoiding flag-planting:

Distinguishing Confirmatory from Exploratory The first apparent advantage to preregistration—often said to be the most important one (Nosek

[1] The first registries were established by medical researchers in the 1960s and were originally designed to help experimenters recruit participants for clinical trials, but as pointed out by Wiseman et al. (2019), preregistration, as we think of it today, started in parapsychology. In 1974, Martin Johnson, a professor of parapsychology and an editor of newly established European Journal of Parapsychology, introduced a preregistration practice for this journal (Johnson, 1975), in an effort to make parapsychology protocols more rigorous. In the editorial, Martin Johnson describes how according to the philosophy of the proposed preregistration model, experimenters should define their problems, formulate their hypotheses and outline their experiments, prior to commencing their studies. In Declaration of Helsinki §19, the World Medical Association (2013) demands: "Each clinical study must be registered in a publicly accessible database before the first test subject is recruited." While the European Commission refers to it, it has not been universally adopted (Rid and Schmidt, 2010).

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 83–93

et al., 2018)—is that it clarifies what counts as confirmatory research, which *has to* preregister, and what counts as exploratory research with no obligation to preregister. Confirmatory research is hypothesis testing, held to the highest standard and which aims to minimize false positives. Here, p-values are generally assumed to have diagnostic value and inferences can be drawn to wider populations. Exploratory research, in contrast, has a different status: It generates rather than tests hypotheses and results should be replicated and confirmed at a later stage. Typically, the focus is on minimizing false negatives, and p-values are not assumed to have diagnostic value (Schwab and Held, 2020). Moreover, findings are not assumed to be directly transferable to wider populations. Rubin (2020), however, points out how it is not always trivial to distinguish between confirmatory and exploratory research: if a researcher, for example, retries a hypothesis from previously published literature to explain an experiment they just ran, is this an *a priori* or a *post-hoc* hypothesis? See also §3.

Fishing Expeditions and Harking Preregistration is often said to prevent *fishing expeditions* and so-called *harking*[2] (Andrade, 2021), namely, post-hoc characterization of hypotheses based on experimental outcomes. Fishing expeditions is ambiguous in the literature (between fishing and harking), but we use the term to refer to cherry-picking dataset and protocols to validate a hypothesis. Harking, in turn, is what researchers do when they indiscriminately examine associations between different variables, not with the intention of testing *a priori* hypotheses but simply hoping to find something of significance. Rubin (2020) calls this 'undisclosed hypothesizing after the results are known.' Having authors preregister their hypotheses potentially improves the reliability of confirmatory research by controlling for cherry-picking and multiple hypothesis testing, implicit to exploratory research. See also §4.

Publication Bias Van Miltenburg et al. (2021) say that, to them, the main advantage of registered reports is that they provide a means to avoid publication bias. Because studies are evaluated *prior* to the results, negative results have the same chance to be published as positive ones. Rubin (2020) refer to this as avoiding the suppression of *a priori* hypotheses that yield null or disconfirming results. Publication bias is claimed to be a serious problem in NLP research by many (Plank et al., 2014; Card et al., 2020; Cohen et al., 2021). The argument was also used in epidemiology, but received some pushback (Loder et al., 2010). See also §5.

Flag-planting Van Miltenburg et al. (2021) also suggest preregistration can prevent so-called *flag-planting*. Flag-planting refers to rushing to be the first to publish results. Flag-planting potentially comes at the cost of scientific integrity and quality. Because of biases in peer-reviewing, it is harder to publish a corrected version of a study that is already out there, than to publish an error-prone study that is the first of its kind. See also §6.

Other Reasons to Preregister We have covered the main reasons van Miltenburg et al. (2021) had for adopting preregistration and will now move to our two-sided, dialogical discussion of its pros and cons. In our dialogue, we will let Zeny and Socart,[3] our house philosophers, debate preregistration. In §3–§8, we will let them discuss arguments *against* preregistration, including arguments that run counter to those presented by van Miltenburg et al. (2021), but we will first let Zeny provide us with a fifth argument *for* preregistration:

ZENY: Socart, there's an additional argument for preregistration, I believe. Early feedback on experimental methodology through a peer-reviewed registration process should improve the quality of the methodology, should it not? Such feedback also saves resources otherwise spent on failed or misleading experiments.

SOCART: Zeny, we both know turn-around is fast in NLP research. Experiments are easier to run and feedback is much faster than for clinical trials, where preregistration is common.

ZENY: NLP as a field has many virtues, but the reviewing cycle is slowing as the field grows larger. Moreover, experiments are becoming more expensive with larger models, creating barriers of entry (Bender et al., 2021) and experiments have substantial environmental impact.

SOCART: You make an important point, Zeny, but early feedback would require more time from reviewers. Since reviewers and researchers coincide,

[2]Short-hand for "hypothesizing after the results are known". Often conflated with fishing, but the two differ: Harking fixes the experiment, varies the hypothesis, so to speak, whereas fishing fixes the hypothesis, varies the experiment. The acronym was coined by social psychologist Norbert Kerr.

[3]Zeny is a mix of Kenny from South Park and Zeno. In Plato's *Parmenides*, Zeno argues for monism—the idea that reality is one stable thing. Socart is a mix of South Park's Eric Cartman and Socrates, who countered this idea by asserting a more nuanced ontology in which things stand in complex relations to each other. Socrates, in other words, took a more nuanced stance, arguing against the existence of a one-size-fits-all hypothesis. Zeny and Socart adopt similar positions in our dialogue about preregistration.

preregistration would potentially save compute resources, but not working hours.

ZENY: That is an oversimplification. Giving feedback on an early draft takes much less time than writing a full paper. If the reviewers are carried over, they will save time when reading the full submission, also. Preregistration would also prevent cherry-picking and invalid use of significance tests by excluding explorations from confirmatory research.

SOCART: But the explorations could be done prior to preregistration and researchers may then be more inclined not to report such explorations at all.[4]

ZENY: Any system can be tricked, but if researchers adopted the practice of preregistration, we would, all things being equal, increase transparency and decrease bias around research.

SOCART: Dear Zeny, you too have seen the evasiveness of bureaucracy, e.g., in NLP conference submission forms. While preregistration reports would initially be light-weight, transparency could easily be clouded by the complexity of assembling the information required for preregistration as new requirements are added over time.[5]

See Bracken (2011) and Rubin (2020) for a discussion of more advantages to preregistration. In addition to reducing fishing expeditions and harking, flag-planting and publication bias, these include: **a)** preventing p-hacking, **b)** prespecifying tolerated significance levels, **c)** identifying selective reporting,[6] and **d)** preventing *forking paths* practice.[7] None of these points are uncontroversial and Rubin (2020) also presents counter-arguments against **a)-d)**. For example, prespecified significance levels have been superseded by the practice of simply reporting *actual* α-levels. Surprisingly, there has been little work on whether preregistration increases trust in science, except for the study by Field et al. (2020), which was under-powered.

We focus on *preregistration for NLP research*. In general, there is no *a priori* reason to think that the pros and cons of preregistration transfer from clinical trials over epidemiology to NLP research. In clinical trials, for example, it is easy to decide when a protocol must be registered. This simply happens before the first subject is assigned to treatment. In epidemiology, there is no such bright line (Lash and Vandenbroucke, 2012) and it is equally hard to see one in NLP. While general machine learning has seen many related methodological discussions (Gencoglu et al., 2019; Lipton and Steinhardt, 2018; Gundersen et al., 2022), there has, to the best of our knowledge, been no published discussions of preregistration practice in this field, with the exception of Gundersen (2021).[8][9]

In our discussion below, we will ignore the most trivial challenges to preregistration, such as deviations from data collection plans for practical reasons, discovery of assumption violations, etc. Such challenges have already been discussed in the clinical literature, e.g., by Nosek et al. (2018).

3 Encouraging Confirmatory Research

We let SOCART and ZENY discuss whether preregistration will succeed in distinguishing between confirmatory and exploratory research. A decade ago, when preregistration was being implemented and discussed in epidemiology, the worry that preregistration would introduce a bias against "the end of the research spectrum that constitutes the quirky, brilliant work that is not enterprise-driven" (Sørensen and Rothman, 2010), was the main concern among its opponents. SOCART and ZENY discuss the consequences of insisting on a distinction that is not trivial to uphold in practice.

SOCART: It seems to me, dear Zeny, that many NLP projects are driven *not* by an explicit hypothesis, but by a desire to understand the behavior of a model, to be able to characterize its strengths and weaknesses, or by a simple gut feeling that at the locus of interacting variables, interesting dynamics can be observed.

ZENY: Can you provide me with an example?

SOCART: Certainly. Pires et al. (2019), e.g., showed that knowledge encoded in multilingual BERT (Devlin et al., 2019), could be transferred across languages—even across scripts, that such transfer worked best between typologically similar languages, that it could process code-switching and find translation pairs. They also showed systematic deficiencies affecting some language pairs. How would they have foreseen these findings? Or even the dimensions that turned out to be of interest? Even if they had foreseen how they wanted to explore transfer across scripts and typological classes, what if genealogy or demography turned out to be more interesting than typology?[10]

[4] This point was also made for preregistration in epidemiology by Sørensen and Rothman (2010).

[5] See Loder et al. (2010) for arguments from epidemiology.

[6] Selective reporting is regarded the most important contributor to irreproducibility by Baker (2016). Nosek et al. (2018), advocating for preregistration, presents similar arguments.

[7] This practice refers to when researchers make decisions about which correlation tests to conduct based on properties of their data. The practice is named after *The Garden of Forking Paths*, a 1941 short story by Jorge Luis Borges.

[8] Workshops on preregistration at ICCV 2019 (https://preregister.vision/) and at NeurIPS 2021 (https://preregister.science/) seemingly did not lead to publications or a change in practice yet, but the website for the 2021 workshop says papers are forthcoming.

[9] Gundersen (2021) complains no AI venues support preregistration, but provide no arguments for or against it.

[10] Independent language families may share features, i.e., be typologically close, but genealogically apart. See Rama and Kolachina (2012) for discussion.

ZENY: I am unconvinced that preregistration would be a serious obstacle to such work. Pires et al. (2019) could have defined the search space in advance – or maybe this is exploratory work that would not have to register in the first place? Remember also, Socart, that the preregistered plan can be updated and refined in the course of a research project. Plans can be revised, but this does not cancel out the benefits of planning.

SOCART: Preregistration may accommodate deviation from the plan, but would risk losing its benefit if researchers were allowed to preregister too many hypotheses or update their plans too frequently. Let us illustrate this with another example. Zhao and Bethard (2020) study how BERT models' learned self-attention functions change during fine-tuning to reflect the target task. They find this to be the case only in smaller models; with more parameters, the change disappears. Imagine now that their hypothesis was confirmed only for select combinations of positional encodings, regularizers and optimizers.

ZENY: This sounds suspiciously like a case of forking paths. Dror et al. (2017) warned us about this risk, encouraging us to at least validate our hypotheses on multiple datasets to reduce the chance of p-hacking.[11] Again, preregistration would not be required for all research.

SOCART: So if authors submitted exploratory work for peer review, would reviewers then decide if bypassing preregistration was appropriate?

ZENY: Yes. Preregistration clarifies the distinction between exploratory and confirmatory research.

SOCART: But what if Pires et al. (2019) had pointed to earlier work already hypothesizing that transfer works best between typologically similar languages? Would this not have made their research confirmatory in the eyes of their readers and therefore in need of preregistration?[12]

ZENY: It very well might have. If they consider it exploratory, they should also point to alternative hypotheses that would explain different results.

SOCART: Moreover, if preregistration becomes a badge of honor or increases your chances of getting your work accepted, because the findings have a different air of trustworthiness,[13] would this not be a reason to encourage your students to perform confirmatory rather than exploratory research? Preregistration would, in other words, inject a bias toward confirmatory research into NLP.

ZENY: I, for one, would welcome this kind of bias.

✓ *Preregistration is challenged by r-harking and may bias NLP toward confirmatory research.*

4 Some Expeditions May Prevent Others

If you ask an NLP researcher if they are "on a fishing expedition" or if they are hypothesizing after the fact, you will instantly make them feel very uncomfortable. It is widely accepted that fishing and harking are bad practices.[14] Socart, however, has an argument *for* (occasional) harking:

SOCART: Two researchers, Ann and Bob, have the same hunch, that the regularization technique R_2 is better than its competitors, R_0, R_1, R_3. Ann realizes after a set of experiments of datasets D_0, D_1, D_2 that, in fact, R_1 is better than R_2. Since this was previously unknown to the community, she publishes it, presenting it (somewhat vaguely) as a confirmation of an *a priori* hypothesis. Bob tests the same hypothesis, i.e., that R_2 is superior to R_0, R_1, R_3. Seeing R_1 is better on D_0, D_1, D_2, he looks for more datasets, until he has a suite of datasets D_4, D_5, D_6 on which R_2 is better than R_1.

ZENY: Bob's cherry-picking is extremely problematic, but so is Ann's harking.

SOCART: But would you agree that granting her the freedom to hark most likely reduces her temptation to cherry-pick?

ZENY: This would simply reclassify Ann's work as exploratory rather than confirmatory. I see no reason why preregistration should not allow this.

SOCART: The two researchers both departed from their original plans, but Ann's willingness to depart from her original hypothesis serves us better than Bob's cherry-picking. In this way, harking can prevent a researcher from taking on a fishing expedition.

✓ *Preregistration should allow for re-classification of confirmatory research as exploratory research.*[15]

5 Solving Publication Bias?

Sørensen and Rothman (2010) argue against preregistration solving publication bias, because researchers can still selectively register studies after preliminary data explorations. Imagine Hippocrates, the Greek physician, was asked to preregister his vivisection experiments. If Hippocrates was studying 10 soldiers with brain lesions, what would prevent him from using one soldier to generate hypotheses, preregister those and conduct the final experiments on the remaining nine? Or worse, peek at all, preregister and go back to the data?

SOCART: Say Hippocrates has two hypotheses about the soldiers, such as that the heart is the seat of

[11] See Belz et al. (2021) for a similar discussion.

[12] This practice is known as "retrieving hypotheses after the results are known" (r-harking) (Rubin, 2017).

[13] Greater reliance on preregistration improves estimation of effect sizes, as shown by Strømland (2019).

[14] Andrade (2021) notes that fishing expeditions can be "ethical" if acknowledged as such, and if appropriate corrections are performed when computing significance results.

[15] Reclassification flags work as exploratory, thereby increasing transparency, but would not impact acceptance decisions. An alternative, suggested by one of our reviewers, would be to introduce intermediate reports as a required step to share the results of the preregistered study before continuing to preregister and test alternative hypotheses a part of the same study. This further increases transparency, and prevents having 'unwanted' results 'swept under the rug' in the final publication. Researchers working on a similar topic would already benefit from the results in an intermediate report.

compassion and that the brain is the seat of rational thought. Upon preliminary exploration, he sees many soldiers have turned cold-hearted by the atrocities of war, but few complain of heartaches. Nearly all soldiers who are delusional or suffer from memory loss, also suffered blows to their heads. Hippocrates pursues and preregisters only the hypothesis that the brain is the seat of rational thought. He has now preregistered, not a prediction, but a post-diction, ignoring the negative result.

ZENY: But Socart, did you not, a moment ago, argue that preregistration would dampen the creativity of research by preventing fishing expeditions and harking?

SOCART: In theory, yes. Preregistration will dampen creativity if properly sanctioned, but I am skeptical that this would be practically possible, rendering preregistration ineffective, an unnecessary administrative burden for all—and a bottleneck for the honest few.

ZENY: If preregistration *prior* to data collection is encouraged, this would solve the problem, no?[16]

SOCART: Surely, but this would mean only one preregistered study per dataset. Since few NLP papers introduce new datasets, this would render preregistration ineffective for the vast majority of NLP research.

ZENY: This is a good point, Socart, but community-wide overfitting to benchmarks is a vice, not a virtue. If preregistration encourages the introduction of new test datasets, that's a good thing, no? Some even argue that all papers should ideally introduce new test data.

To the contrary? ...in which SOCART and ZENY continue to discuss whether preregistration could actually make publication bias worse. SOCART suggests that preregistration could amplify publication bias, if positive results are still preferred over negative ones and preregistration forces researchers to focus on predictably positive results, arguably a small subset of the positive results. If papers are accepted on the basis of preregistration, this could increase an arguably already existing bias toward incremental improvements.[17]

SOCART: You say preregistration will make it easier to publish negative results, because studies are evaluated prior to obtaining results?

ZENY: That is correct, Socart.

SOCART: ...but do we really know why there are so few NLP papers about negative results? See, in NLP, negative results are much harder to establish than positives. If I want to show that self-attention or weight averaging does *not* lead to improvements for some problem, I need to show that holds across all implementations, all architectures and all available datasets. The value of a report stating that for one such combination, self-attention didn't do much, would be next to nothing. Isn't that the real explanation for the skew in the NLP literature?

ZENY: Negative results are key to scientific progress (Barwich, 2019), but are hard to establish if they are very general. Published positive results often overclaim their generality. Both should be confirmed only by accumulated evidence in diverse settings.[18]

SOCART: No, no, you fail to see there's a qualitative difference, Zeny! Imagine if Ann was evaluating self-attention for sentiment analysis. To answer the hypothesis that self-attention works in the positive, she just needs a significant result in a single setting. In contrast, in order to establish a negative result, she has to explore *all* possible settings.[19] How would preregistration make establishing a negative result less formidable a challenge?

ZENY: I agree that it would not. My only claim is that evaluating studies prior to obtaining results would prevent any bias on behalf of peer reviewers to evaluate negative results more harshly.

SOCART: ...but in reality, we do not know if such a bias exists, or whether it is only fair that such a bias exists, because the bar by definition should be higher for negative results?

✓ *Preregistration may increase publication bias.*

6 Solving Flag-Planting?

Flag-planting is one of the motivations for preregistration for van Miltenburg et al. (2021), but *exclusive* preregistration may also, conceivably, have the opposite effect.

SOCART: Say Ann and Bob get the same great idea—e.g., to evaluate the sensitivity of textual entailment models to presupposition projection—and worry that they will be scooped before getting around to publishing it. Ann and Bob now follow two different strategies: Bob rushes to preregister a study hypothesizing that state-of-the-art models are sensitive to such phenomena, while Ann rushes to run the experiments and publish the paper. Zeny, which strategy is better for science?

ZENY: I would say it's Bob's, since rushed experiments are more likely to be flawed.

SOCART: But Bob plants his flag faster than Ann, essentially scooping her. By doing so, Bob discourages Ann from pursuing this idea by planting his flag first. What if Bob fails to conduct proper

[16] This is implicit in preregistration in scientific fields where data is not re-used, e.g., in psychology (Wiseman et al., 2019).

[17] Lash and Vandenbroucke (2012), for example, argue that: "prespecified hypotheses often take little risk, invoke little imagination and stray only a short distance from what is already well understood."

[18] NLP has seen relatively few meta-studies (Cramer, 2008; Søgaard, 2013; Hoyle et al., 2021; Bugliarello et al., 2021), but hopefully, we will see more in the future.

[19] We flesh this out a bit. The research hypothesis in Ann's case is that self-attention helps. What this means is that in *some* implementation, it leads to robust improvements. The vast majority of NLP hypotheses take this form: X can, in some implementation, lead to general improvements on one or more tasks. If the baseline is fixed, this amounts to existential quantification ("some"). Conversely, its negation ("self-attention does not help for sentiment analysis") amounts to universal quantification, i.e., there is *no* implementation in which this is the case. This is obviously much harder to prove than the corresponding positive result.

experiments altogether? Had it not been for preregistration, both researchers would have pursued their idea, providing mutual replication and increasing the likelihood of the idea materializing into an actual result.

ZENY: Some would say this is one of the advantages of preregistration: Ann pursuing the same idea would have been a waste of time.

SOCART: This assumes Ann and Bob would have conducted their research in exactly the same way and that none of them were prone to error. In other words, that researchers are machines that simply execute their unambiguous experimental protocols. I think preregistration just moves flag-planting to earlier in the research process, lowering the bar for researchers to plant their flags, since less work is required to plant a flag. And when a bar is lowered, more researchers are likely to plant more flags.

ZENY: Van Miltenburg et al. (2021) explicitly encourage concurrent work.

SOCART: Yes, I did read that passage, but they do not discuss *how* preregistration would impact concurrent work. Do they envisage a review system in which Ann is allowed to follow up on the idea that Bob preregistered?

ZENY: I'd do that, in the spirit of open science.

SOCART: Such *inclusive* preregistration would clearly discourage protectionist researchers from preregistering their studies. If a preregistered study is up for grabs for other research labs, labs with more resources could likely wrap up the experiments faster than the researchers who registered it.

ZENY: . . . unless we envisage a review system allowing Ann to preregister the same study, giving Ann and Bob equal chances to pursue the study.

SOCART: This would be equivalent to telling reviewers of a paper to consider as "concurrent" any other work published within the last 1–2 years (assuming this is the approximate life span of a research project), including preregistered studies. Today, reviewers are told to disregard work published within the last three months, but already, reviewers seem to ignore this guideline in practice, presumably because they do not want to compromise the fast turn-around in NLP research.

ZENY: But shouldn't we incentivize slow science? Many NLP papers neglect related work and keep reinventing the wheel. We need deeper analysis to enable disruptive scholarship and novel ideas.[20]

SOCART: Slow science also has disadvantages. Fast turn-around has had many positive effects on NLP, including rapid replication. Projects can become "too big to fail," causing confirmation bias.[21] Lowering false positive rates is important, but so is healthy distrust in published results.

✓ *Preregistration may increase flag-planting.*[22]

[20] Chu and Evans (2021) showed that fast turn-around results in stymied fundamental progress in large scientific fields.

[21] This can result from financial interests (Ioannidis, 2005), e.g., due to "sunk cost" (Perignat and Fleming, 2022).

[22] One obvious solution is to make preregistration non-public, but then preregistration would not prevent two groups doing the same study.

7 Solving p-Hacking?

Inflation bias, also known as p-hacking, refers to selective reporting to produce statistically significant results. Søgaard et al. (2014) lists several p-hacking techniques used, perhaps inadvertently, in NLP papers. If a statistically significant result is seen as the key to getting your paper accepted, researchers are presumably willing to go far to squeeze out a small p-value. But if preregistration facilitates the publication of negative results, it seems it would also reduce the incentive to engage in so-called p-hacking, e.g., obsessive fiddling with data and models until reaching the magical $p < 0.01$. It has been noted, however, that preregistration leaves plenty of room for p-hacking (Bakker et al., 2020). Generally, eliminating p-hacking entirely is unlikely when career advancement is assessed by publication output, and positive results are favored by scientific peers (Head et al., 2015).

Socart and Zeny discuss whether preregistration will reduce or amplify the incentive to engage in p-hacking:

SOCART: Imagine Ann again, who is now evaluating if self-attention is helpful for sentiment analysis. Say she preregisters the hypothesis that self-attention *is* helpful, only to find that her first results are negative. We would now like Ann to go ahead and acknowledge the negative results on print, right? However, as we just saw, when your first results are negative, more results are typically needed to draw a firm, negative conclusion that self-attention does *not* help. Sometimes more data collection is needed and more human evaluations may be needed. Pursuing the negative result will, in other words, be a lot of work.

ZENY: But very important!

SOCART: Preregistration increases the amount of work that goes into moving your focus to establishing a negative result: You will need to augment your preregistration with information about the experiment, your new hypothesis and the new experiment you plan to perform.

ZENY: Documentation has to be light-weight.

SOCART: . . . but preregistration would get people more invested in their ideas and bias them in how results are interpreted. When people go on record with a study description, they will defend why it's reasonable and likely leading to a positive result. Researchers are always prone to confirmation biases, but now social expectations and reputation will amplify their existing biases. This would lead to the opposite of the effect intended.

✓ *Preregistration may increase p-hacking.*

8 Risk Tolerance

Attempts to reduce false positives tend to also lead to reductions in true positives. Many applications

require near-zero false positive rates, but most NLP experiments show low risk of direct negative impact on society or individuals therein,[23] as indicated by the relatively few papers receiving ethical reviews. Hence, we can afford to take risks and explore hypotheses that end up wrong. Parascandola (2010) reminds us how this is a key ingredient in increasing knowledge and reducing uncertainty, getting us off the beaten track. NLP benefits from being frequently wrong and implementing preregistration to prevent false positives has a drawback. In §9, we will argue that what is important is to balance preregistration with our risk tolerance.

SOCART: Imagine Ann works on hate speech detection for a social media company. Bob works on topic classification of social media posts at the same company. They both validate and evaluate the models in the wild on beta users. They both can use logistic regression and SVMs. SVM is sometimes superior, but exhibits more variance across hyper-parameters. If I were to advise Ann or Bob to use logistic regression, who then?

ZENY: Probably Ann, since we can tolerate less risk in her situation. But how would preregistration affect the risk tolerance of researchers?

SOCART: Imagine if you were asked today to carry a solid bottle of olive oil over to Plato's house and tomorrow to bring him a fragile, beautiful vase decorated with gold. On which of the two days would you be more inclined to *run* there?

ZENY: Today, but how is this relevant? What are the tasks where we can afford to 'run'?

SOCART: For tasks in which false positives are associated with high risk, we should hedge our bets by preregistering conservative hypotheses; for other tasks, this sort of inhibition of is unfortunate.

ZENY: This is exactly why exploratory research still has a place in a world with preregistration—namely, for tasks where we can tolerate risk.

One may argue that risk mitigation is not what preregistration is for. The purpose of institutional review boards (IRBs) and ethics reviewing is to flag and prevent too risky studies (IRBs focus on risk to human participants, while ethics reviewing also addresses potential applications). We have three reasons to think preregistration requirements should depend on expected risk: (a) It is impossible to review the implications of a study before you have a solid study plan. If preregistration includes risk assessment, this could provide input for IRBs and ethics reviewing (or, in a more distant future, be part of the same process). (b) A partial roll-out of preregistration may help us balance Type 1 and Type 2 errors. Expected risk affects the cost of false positives and hence the optimal balance between Type 1 and Type 2 errors. Since bureaucracy, by the end of the day, also incurs a cost on society, this reinforces our belief that mechanisms should be implemented only for where there is direct impact on society at large. (c) Finally, ethics reviewing is typically part of the standard review process, i.e. *after the fact* and can therefore not respond to malpractice in the experimental design or prevent publication of preprints.

Overly cautious preregistration practice may, in sum, decrease our true positive rate and add bureaucratic overhead to research practices without proper motivation. An all-over-the-map roll-out of preregistration would change the risk tolerance in research and society, just like registration and documentation has increased risk sensitivity in the past. Simultaneously, evaluating risk early on has clear advantages over the current review process.

✓ *Preregistration may lower our risk tolerance.*

9 A Proposal

We have tried to present pros and cons of preregistration. If we have focused a bit more on the cons, this is only because van Miltenburg et al. (2021) did a great job highlighting its advantages. We will, if anything, argue for only a partial roll-out of preregistration of NLP research. Preregistration is a way to minimize harms of NLP research, but only when risk is high. To motivate this, consider, as first noted by Lash and Vandenbroucke (2012), two seemingly opposed arguments *for* preregistration: a) Preregistration counters the suppression of (negative) results. b) Preregistration identifies false positives. Lash and Vandenbroucke (2012) argue that while (b) is a valid argument for preregistration of *clinical trials*, it is not a valid argument for preregistration in the context of mere "accumulation of evidence" (Lash and Vandenbroucke, 2012). Here, concerns about balancing Type 1 and Type 2 errors disappear. Preregistration mitigates risks associated with research, reducing potential harms, but at the cost of scientific progress. This calls for a cost-benefit analysis: How much risk can be tolerated for what potential gains?

One way to frame this discussion in NLP is to

[23] This does not refer to the downstream risks after deployment, just the risks associated with the research experiments. Two reasons for NLP experiments being relatively low risk are the rare involvement of human participants in NLP experiments and the historical focus on professionally generated text (Hovy and Spruit, 2016). We are seeing a shift toward human-in-the-loop evaluations and user-generated content, but this still makes for a small fraction of NLP research.

ask how afraid we should be of being wrong. In clinical trials, there is a significant cost to being wrong. In biomedical studies, false positive rates have been found to be around 14% (Jager and Leek, 2014). Whatever the number is for NLP, lowering it by adding more checks, will lead to a drop in the *true* positive rate. If a false positive could result in human tragedy, the price of a lower true positive rate is worth paying, but in NLP, the cost of a false positive is often paid for in compute and human hours. While both can be scarce resources, the open access nature of NLP makes being wrong less dangerous, since mistakes are quickly corrected.[24] Zeny would object that pretraining of language models is not easily reproducible (Bender et al., 2021). Pretraining very large language models should maybe be required to preregister and this would possibly require revising the questionnaires provided by van Miltenburg et al. (2021). Another concern is how NLP contributes to social and cultural inequality (Hershcovich et al., 2022). If NLP research is likely to help some more than others, this may be reason to require preregistration. Here, the questionnaires provided by van Miltenburg et al. (2021) would also be insufficient.[25]

So what we propose here reflects a middle-of-the-road position on preregistration. The idea is to **limit preregistration to research for which our risk tolerance is low**. This prevents most of the adverse effects of preregistration, e.g., publication bias, flag-planting and p-hacking. NLP research is subject to IRB and ethics reviewing, but we believe this should be merged with preregistration (§8). Reports should be reviewed and reviewers follow the submission (§2).[26] Currently, we accept and reject papers through blind peer-reviewing, but some papers are accepted *conditional* on a positive ethics review. We propose a reviewing procedure in which some work is only accepted conditional on the work having already been registered with positive reviews. For researchers, this would mean you need to get your preregistered reports accepted, before you initiate the research project. Once completed you will send the final submission in for a new set of reviews, hopefully by the same reviewers. This procedure is somewhat cumbersome and has all the disadvantages we discussed above. Therefore, it should only be used when it is deemed necessary, i.e., when the expected risk of the NLP research is sufficiently high.

A paper which a) is confirmatory and b) concerns an application for which risk tolerance is low, can be rejected for not having preregistered. We noted the necessity of allowing preregistered research to re-classify as exploratory, i.e., conditional acceptance for non-preregistered, flagged research, if it explicitly labels itself as exploratory. This would lead to three different categories of accepted papers: (i) non-preregistered (confirmatory or exploratory) research, (ii) preregistered, confirmatory research and (iii) non-preregistered research *marked explicitly as exploratory*. This would, we argue, give us the advantages of preregistration where they are most needed, e.g., where false positives are associated with very high risk.

We left one important thing in the open for now: How do we fairly decide if a research subject and protocol warrants low risk-tolerance? Ethics review board members are already asked to flag work that 'exhibits an increased risk of harm outside the current norms of NLP or CL research'.[27] This can be hard to determine, but board members *already* have to make this difficult decision. Ethics reviews could learn from established risk assessment frameworks (Schwerdtner et al., 2020).

10 Conclusion

Our two-sided dialogue has discussed pros and cons of preregistration in NLP, building on similar discussions in epidemiology. What opponents elsewhere have proposed as alternatives to preregistration is already found in NLP research: open access, common repositories and data sheets (Lash and Vandenbroucke, 2012). Preregistration, we argue, is less urgently needed in fields that already facilitate **replication** and where **risk of false positives** is

[24] Of course researchers are sometimes blind-sighted by scientific paradigms and hype, biasing their interpretation of results. Such dynamics is central to, e.g., the Popper-Kuhn debate (Rowbottom, 2011), but beyond the scope of this paper.

[25] Specifically, we think the questionnaire for 'NLP Engineering experiment paper' (§A.3) should include questions about computational resources needed for pretraining. Since the risk of wasting resources is high for language model retraining, preregistration and early feedback may be particularly useful for such research, but the review of the registered point would have to take this information into account. To mitigate social and cultural inequality, we propose to revise the questionnaire for 'Resource paper' (§A.6), adding questions about the demographics of data sources and annotators, as well as making a corresponding explication of social and cultural concerns in Question 11 of §A.3.

[26] This would be hard to coordinate for most fields, but in NLP, the ACL Rolling Review platform could make it easier.

[27] https://aclrollingreview.org/ethicsreviewertutorial

low. The **fast turn-around** of NLP research means the advantages of transparency and early feedback are smaller. Nevertheless, society's risk-tolerance varies across NLP applications. Legal or medical decision support systems are high-risk application areas. Here, we need to consider all safety measures on the table, including preregistration.

Impact Statement

Preregistration is one of several practices that promote responsible, high-quality research. Others include replication, transparency and open access, as well as impact statements and explicit discussion of study limitations. All such practices come with pros and cons and it is key to scientific progress and positive impact that scientific communities evaluate which practices are adequate in their domain. The increasing real-world impact that NLP research has exhibited recently and will likely continue to exhibit warrants a careful reconsideration of which practices are called for. Since a major driver of the same impact is the fast-paced exploratory research that characterizes the field, limiting such research may have negative effects as well (see §6). We therefore believe our two-sided debate will enable an overall better outcome in terms of impact.

Limitations

Our discussion of preregistration is inspired by discussions in epidemiology. Many of the concerns epidemiologists had with preregistration seem more relevant to NLP research than the considerations that, by and large, led clinical research to adopt preregistration as a mandatory practice. While we present a proposal for how to implement preregistration in NLP in §9–a proposal that differs from the one presented by van Miltenburg et al. (2021)–our main contribution is a two-sided discussion of its pros and cons, leaving many questions in the air. Our paper is intended to get the preregistration debate off ground, not to nail it to the floor.

Acknowledgements

Thanks to Asbjørn Hróbjartsson and Klemens Kappel for providing us with a helpful overview of the literature on preregistration in epidemiology. Thanks to our anonymous reviewers, as well as all of CoAStaL, for useful feedback on the above discussion. Anders Søgaard received financial support from Innovation Fund Denmark, Google Focused Research Awards, and the Novo Nordisk Foundation.

References

Chittaranjan Andrade. 2021. Harking, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *The Journal of clinical psychiatry*, 82.

Monya Baker. 2016. Is there a reproducibility crisis? *Nature*, 533:452–454.

Marjan Bakker, Coosje Veldkamp, Marcel Assen, Elise Crompvoets, How Hwee Ong, Brian Nosek, Courtney Soderberg, David Mellor, and Jelte Wicherts. 2020. Ensuring the quality and specificity of preregistrations. *PLOS Biology*, 18:e3000937.

Ann-Sophie Barwich. 2019. The value of failure in science: The story of grandmother cells in neuroscience. *Frontiers in Neuroscience*, 13.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Eric Bogert, Aaron Schecter, and Rick Watson. 2021. Preregistration of information systems research. *Communications of the Association for Information Systems*, 49.

Michael Bracken. 2011. Preregistration of epidemiology protocols: a commentary in support. *Epidemiology*, 22:447.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Chris Chambers. 2019. What's next for registered reports? *Nature*, 573:187–189.

Johan S. G. Chu and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118.

Kevin Cohen, Karën Fort, Margot Mieskes, Aurélie Névéol, and Anna Rogers. 2021. Reviewing natural language processing research. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 14–16, online. Association for Computational Linguistics.

Irene Cramer. 2008. How well do semantic relatedness measures perform? a meta-study. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 59–70. College Publications.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Sarahanne M. Field, E.-J. Wagenmakers, Henk A. L. Kiers, Rink Hoekstra, Anja F. Ernst, and Don van Ravenzwaaij. 2020. The effect of preregistration on trust in empirical research findings: Results of a registered report. *Royal Society Open Science*, 7(4).

Oguzhan Gencoglu, Mark van Gils, Esin Guldogan, Chamin Morikawa, Mehmet Süzen, Mathias Gruber, Jussi Leinonen, and Heikki Huttunen. 2019. Hark side of deep learning – from grad student descent to automated machine learning.

Odd Erik Gundersen. 2021. The case against registered reports. *AI Magazine*, 42(1):88–92.

Odd Erik Gundersen, Kevin Coakley, and Christine Kirkpatrick. 2022. Sources of irreproducibility in machine learning: A review.

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3):1–15.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan L. Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2018–2033.

John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Med*, 2(8):e124.

Leah Ruth Jager and Jeffrey T. Leek. 2014. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15 1:1–12.

Martin Johnson. 1975. Models of control and control of bias. *European Journal of Parapsychology*, 1:36–44.

Timothy L Lash and Jan P Vandenbroucke. 2012. Should preregistration of epidemiologic study protocols become compulsory?: Reflections and a counterproposal. *Epidemiology*, 23(2):184–8.

Zachary C. Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship.

Elizabeth Loder, Trish Groves, and Domhnall Macauley. 2010. Registration of observational studies. *BMJ (Clinical research ed.)*, 340:c950.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.

Mark Parascandola. 2010. Epistemic risk: empirical science and the fear of being wrong. *Law, Probability and Risk*, 9(3-4):201–214.

Elaine Perignat and Fraser F. Fleming. 2022. Sunk-cost bias and knowing when to terminate a research project. *Angewandte Chemie International Edition*, 61(36):e202208429.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar. Association for Computational Linguistics.

Taraka Rama and Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of COLING 2012: Posters*, pages 975–984, Mumbai, India. The COLING 2012 Organizing Committee.

Annette Rid and Harald Schmidt. 2010. The 2008 declaration of helsinki — first among equals in research ethics? *Journal of Law, Medicine & Ethics*, 38(1):143–148.

Darrell Rowbottom. 2011. Kuhn vs. popper on criticism and dogmatism in science: A resolution at the group level. *Studies In History and Philosophy of Science Part A*, 42:117–124.

Mark Rubin. 2017. When does harking hurt? identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, 21:308–320.

Mark Rubin. 2020. Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, 16(4):376–390.

Simon Schwab and Leonhard Held. 2020. Different worlds confirmatory versus exploratory research. *Significance*, 17(2):8–9.

Paul Schwerdtner, Florens Greßner, Nikhil Kapoor, Felix Assion, René Sass, Wiebke Günther, Fabian Hüger, and Peter Schlicht. 2020. Risk assessment for machine learning models.

Anders Søgaard. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta, Georgia. Association for Computational Linguistics.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.

Henrik Toft Sørensen and Kenneth J Rothman. 2010. The prognosis for research. *B M J*, 340:c703.

Eirik Strømland. 2019. Preregistration and reproducibility. *Journal of Economic Psychology*, 75:102143. Replications in Economic Psychology and Behavioral Economics.

Dennis Ulmer, Elisa Bassignana, Max Müller-Eberstein, Daniel Varab, Mike Zhang, Christian Hardmeier, and Barbara Plank. 2022. Experimental standards for deep learning research: A natural language processing perspective. In *ML Evaluation Standards Workshop at ICLR 2022*.

Anna Veer and Roger Giner-Sorolla. 2016. Preregistration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology*, 67.

Richard Wiseman, Caroline Watt, and Diana Kornbrot. 2019. Registered reports: An early example and analysis. *PeerJ*, 7:e6232.

World Medical Association. 2013. World medical association declaration of helsinki ethical principles for medical research involving human subjects.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.