Reviewer 1

R1.1: My biggest concern is whether the created texts truly capture what coders were expected to find, since it is striking that also the expert coding of frames works poorly. Mentioning this issue in the discussion is insufficient, and more information is needed to validate your approach. Specifically, the construction of the counterfactual dataset lacks some transparency. Please explain more clearly how it is created, provide examples of resulting articles, and highlight their differences along the two dimensions.

A1.1 Thank you for raising this issue, following this suggestion and the suggestions of Reviewer 2 (R2.1), we now conducted additional analyses to assess what caused the problems for expert coding, and discuss whether there is a problem of authorial intent not being accurately expressed in the articles, or problems with human coders not accurately identifying the intended frame, or whether conceptual issues with Generic Frames are the root cause of our surprisingly low correct classification by expert coders.

We made our construction process more transparent by moving the description of the coding instructions from the appendix to the main paper and explicating the steps the student authors took to assess each other's article to ensure the intended frame was present in the articles (pp. 12-13).

Furthermore, we now conduct additional analyses to identify the causes for low expert coding validity. We identify several possible causes through a second expert coding where we asked the experts to rate how accurately they judge the articles to portray the intended frame: We find that the Morality Frame was often not identified due to a lack of explicit moral messaging, and the implied moral message not being clear enough in the articles, which is in line with research indicating that latent moral messaging is hard to detect (Weber et al, 2018). We also observe that articles often seem to contain additional generic frames to the intended frame. Based on these results, we performed additional sensitivity analyses - once without Morality Frame articles, once with a model that doesn't assume a single frame per article (reported in the Appendix).

We discuss the results in the Discussion section: We note that authorial intent is rarely included in validation checks for frame analyses, which means that several alternative routes are open to expand on our results: Either, we need additional research to validate whether (generic) frame measurements capture authorial intent, or we need to theorize whether generic frames possess enough theoretical exclusivity to allow for frame identification, or we need to theorize how important authorial intent is for a framing model that assumes a causal path from journalist to

media content to audience reception, or we check whether our dataset represents an outlier.

Concerning the tangible validity of the constructed article corpus, however, we argue that it reflects typical patterns of journalistic content production (featuring elements of multiple Generic Frames simultaneously, having issues to explicitly express morality) which increases its external validity. Therefore, the shortcomings identified in these additional explorations of the corpus do not corrupt its value for being used to validate annotation procedures. Yet, they mean that such validation attempts need to be conducted with a view to the specific features of included articles.

To facilitate this process for other researchers interested in working with our corpus, we have not only made the entire corpus but also all the materials available. However, we explained to the AEs at our first submission that we cannot share the OSF link for review (even the anonymous link) because we found that the materials might de-anonymize the submission, even if shared using an anonymous OSF link. The exact reason is that the preregistration must be named (R1.17).

We therefore attached the corpus, the instructions, the anonymized preregistration as well as all the documented we have regarding the project to this revised submission. All of these materials will be available to the research community via OSF upon publication. In the manuscript, we also included more information on how the data was created. As the corpus has been made public, we are not so sure whether it is necessary to include examples in the manuscript text.

R1.2 It is unclear how a synthetic counterfactual dataset can help validate or invalidate a measure meant for real-world data when you state that no such data exists in the real world. Please provide a justification for using a synthetic counterfactual dataset in this context. This might be one of the reasons why even expert coders were not able to identify the right frames in 50% of the cases. Might it be that some combinations just don't work well in real articles?

A1.2 We provided a justification of why topics and frames must be independent in order to have a say about the validity of frame identification methods under the section "Generic frame detection as a test case and ensuring independence from topics". We combined this explanation with our rationale for using Generic Frames as a test case in our synthetic dataset.

Our reasoning is that in real-life data, (sub-)topics and frames often correlate. With automated tools being able to pick up topics, we fear that these methods often pick up frames correctly as an artifact by detecting topic features instead of actual frame features. The same could even be true for human coding: Coders could infer the

presence of a frame not from identifying frame elements, but by recognizing that a topic is prone to correlation with a topic, and correctly guessing the frame from there. Therefore, we assume that to test if a method is genuinely able to pick up frames, we have to ensure that frames are statistically independent from topics. We now clarify that reasoning in the sub-section "Generic frame detection as a test case and ensuring independence from topics" and explicate it again in the following section with the following remark: “This allows us to test whether our frame induction methods are able to find the intended frames without –for human coders-- relying on intuitions about which frames commonly occur within certain topics or –for automated methods-- picking up frames incidentally as a by-product of actually generating distinct topics.”

Randomization is the long established method to generate such independence. For instance, in real-life the media exposure to a certain message is associated with certain demographics. To counter this imbalance, communication researchers have long been using randomization in experiments to balance out the media exposure to messages among different demographics synthetically (as well as in many other fields such as medical science where results from randomized controlled trials are the only accepted standard of drug efficacy). One could refute also that experimental data can never resemble real-life observational data, e.g. media exposure is never random. But it is an effective way (or the “gold standard”, as in medicine) to establish causality. Our case is the same: given a possible alternative explanation (topics as frames), we must use the synthetic approach to establish the direct causality: a method is identifying frames, not topics. On the other hand, given the alternative explanation we can argue that analyzing factual observation data is probably not able to validate a measure.

R1.3 The overall n is small, as you mention in your discussion. I think this could be especially problematic in relation to my previously raised points. If let’s say only 70% of your created articles are a well reflection of the frames they are meant to employ, it really leaves not much room for the tools to work well (the same is true for the expert coders). But again, this might not be a big issue if you can show that the articles are indeed valid examples of the frames they are meant to employ.

A1.3 In response to this comment and comment R1.1 we carried out additional analyses to assess the quality of our original corpus. Using additional expert coding, we find that two possible issues explain the low correct classification rates for human coders: One, the Morality Frame is often communicated very implicitly, and consequently not often picked up by coders. Two, while writing an article communicating one frame, it appears the authors often also deployed elements of other frames, leading to the impression by coders that more than one frame was

present in the document and misidentifying the dominant frame. Combining this insight with comments from Reviewer 2 (R2.1) in the discussion, we debate possible methodological and theoretical implications of this and the need for future research to figure out where the root of the problem lies.

R1.4 For all automated methods you set k=5. I understand that you are aiming to identify 5 different frames, but what room is then left for non-meaningful statistical associations between words that always occur in textual data?

A1.4 Reviewer 2 also raised a similar concern (R2.6). We would like to answer these two concerns jointly. We have attempted to simulate a large k but we decided to give up and acknowledge this as a limitation in the discussion instead, not least because of the 1-mo revision cycle. With this and all other limitations such as our limited n (R1.3), we "downgrade" the benchmark of automatic methods to a preliminary analysis. Instead, we focus more on manual content analysis (also a suggestion by Reviewer 2, R2.1).

Using our current setting, one iteration (one dot) requires 120 (k! = 5! = 120) rounds of computation to find out CCR_{max}. All multiverse analyses took two days on a regular computer. Increasing k and mapping it back to 5 is exhaustive search of partitions. Although it is usually what human coders do, it is very computationally intensive to conduct that exhaustively. The number suggested by R2, for example, is k = 25 and then maps these 25 clusters to 5 frames. The total number of non-empty partitions $s(25, 5)$, i.e. Stirling number of the second kind, is 2436684974110751 (or 2×10^{13} times of 120. Suppose the computational time is constant, it would take 109589041096 years; which is longer than the age of earth). We don't have the computational power to conduct this exhaustive search.

R1.5 The paper would benefit from a clearer explanation of how the gold standard was established. What were the questions used by the RAs to identify the frames? Consider adding these questions to an appendix, as they may be unfamiliar to some readers (in addition to just citing S&V 2000).

A1.5 The coding scheme is now included in the Online Appendix.

R1.6 The "Best case" Correct Classification Rate section needs a brief introduction explaining that you are testing the best case scenario, where the true cluster-to-frame matchings are unknown, and the best performing matchings are used. Also, this section would benefit from a short introductory sentence instead of suddenly starting with a technical sentence like "Suppose y to be…"

A1.6 We added one starting sentence to that paragraph.

R1.7 Please clarify what the null value of CCR mx means and why 89% CIs were used.

A1.7 In the last paragraph of the method section, we explained what the null values of CCR_{max} mean. We responded to R1's query about 89% CI in A1.14.

R1.8 To make the presentation of results more concise, consider combining all figures into one or creating a single table comparing the results of different approaches (e.g. by just stating what the mean improvement over the null is of each method). My reason here is that is somewhat annoying to browse through 6 different figures and combine their information to get to the main conclusion).

A1.8 We took the suggestion by Reviewer 1 and moved all other charts to the Appendix, as they are less useful now.

R1.9 In addition, it would be interesting to see if there is some systematic pattern in which preprocessing steps lead to improved results of the automated methods. Consider using a descriptive specification curve plot (as shown for instance in Simonsohn et al. (2020). This way systematic patterns would be easily identifiable.

A1.9 Thank you very much for this suggestion. After careful consideration, we decided to not implement this suggestion at this stage, as the importance of automatic analyses has been deemphasized in the manuscript in total following the other comments from both Reviewers (see, A1.4).

R1.10 In your evaluation of automated measures, what serves as the ground truth: the "true" data as created in the dataset, or the coding by RAs/experts?

A1.10 We now explicitly mention that the ground truth frame was used in the calculation of CCR_{max} in all cases (see p. 16).

R1.11 Provide a brief overview of each computational method used, either within the main text or as an appendix, if space constraints are an issue, as not all readers might be familiar with these tools.

A1.11 We now provide a short overview over the methods in the Online Appendix.

R1.12 In the gold standard section, please explain the meaning of black and light grey in the figure.

A1.12 The explanations can be found in the figure caption.

R1.13 Provide examples of the created texts and clarify whether creators were instructed to create the same content that RAs/experts were meant to detect.

A1.13 Please refer to A1.1

R1.14 When discussing the confirmation or rejection of hypotheses, restate the hypotheses and explain why 89% credibility intervals were used.

A1.14 We are happy that both R1 and R2 (R2.9) are raising this "why" question on the 89% credibility level. We agree that researchers should alway justify their chosen Type-I rate, although this is barely done (and barely questioned) in some other statistical approaches. In our first submission, we provided a reference, i.e. McElreath. As it turned out, this is not a sufficient justification. It is hardly possible to explain why McElreath chose this level of credibility other than he deliberately chose an arbitrary number that is not 95%. We quote the text in verbatim from McElreath below:

> *The most common interval mass in the natural and social sciences is the 95% interval. This interval leaves 5% of the probability outside, corresponding to a 5% chance of the parameter not lying within the interval. This customary interval also reflects the customary threshold for statistical significance, which is 5% or $p < 0.05$. It is not easy to defend the choice of 95% (5%), outside of pleas to convention. Often, all confidence intervals do is communicate the shape of a distribution. In that case, a series of nested intervals may be more useful than any one interval. For example, why not present 67%, 89%, and 97% intervals, along with the median? Why these values? No reason. They are prime numbers, which makes them easy to remember. And these values avoid 95%, since conventional 95% intervals encourage many readers to conduct unconscious hypothesis tests.*

Given both Reviewer 1 and Reviewer 2 are both interested in *why* the credibility level is 89%, this issue cannot be solved by switching from 89% to the so-called standard of 95% (except we plead to convention, in McElreath's words), as, following McElreath's argument all levels of confidence are naturally equally arbitrary.. Also, we are not philosophers of science and we cannot provide a good justification on why one level of credibility is more defensible than another level. What can be said, however, is that in order to underscore this arbitrary nature of confidence levels employing a 89% HDI has become somewhat of a convention in Bayesian modeling.

To solve this fundamental philosophical question, McElreath also mentions that what really matters is the shape of the distribution of an estimand (100% HDI also with the shape). Therefore, we decided to take McElreath's advice and use the entire distribution instead. It also drives away the perceived awkwardness of using an unconventional credibility level.

R1.15 Please also cite the foundational multiverse papers, such as Steegen et al. (2016) and Simonsohn et al. (2020)

A1.15 We cited Steegen et al. (2016)

R1.16 Lastly, consider using active voice rather than the double passives that appear here and there in the paper for a clearer writing style.

A1.16 Thanks for pointing this out, we went through the manuscript once again, trying to find stylistically more sound solutions for this issue.

R1.17.Finally, please provide your (anonymized) preregistration, as it is not possible to check if the study indeed follows the preregistered protocol without it.

A1.17 See A1.1

R1.18 There is a little typo in your title, it currently reads "[..] for evaluting [..]".

A1.18 Thanks for pointing out this mistake. We fixed it.

R1.19 Regarding the title, and this is just a suggestion, it might be a better fit to already hint at your finding ("computation frame measures don't work") as you are not primarily introducing a dataset in this paper but make the claim that we should be skeptical about computational fame measures.

A1.19 We would like to thank Reviewer 1 for this suggestion. However, after implementing the suggestions and reacting to the concerns raised by both Reviewers, the article is now even more directed towards synthesizing the dataset and discussing its specific features. Therefore we did not follow this suggestion.

Reviewer 2

R2.1 I'm not sure that the paper as currently framed draws as radical a conclusion as it could from the empirical result. The 'gold standard' human coding result is noted to be also poor, albeit better than the automated methods. Arguably, too poor to be relied on as a source of truth in downstream analyses. My own experience of manual frame induction is that different coders will invariably identify different frames, with very poor IRR (and this is made explicit in Burscher et al., 2014, "Teaching the computer to code frames in news" where Krippendorff's alpha of well under 0.5 is reported for the human coders). If this is the case, as the results of this study show, then maybe this should be challenging the whole enterprise of surely the whole enterprise of quantitatively coding frames in texts. The authors are right to draw attention to the difficulties in principle of identifying intention using textual content alone and my own view is that these problems go to the root. If we are to continue to do frame content analysis we should expressly generate a new statement of how we are treating frames to replace Entman's, and be upfront about the extent to which this new conceptualisation is both necessary and significantly different from its predecessors. Given the wide diversity (or incoherence?) of conceptualisations of framing in the existing broader communications literature, this would be following in a grand tradition.

A2.1 We now added a part in the discussion that acknowledges that our results could indicate conceptual problems with the way we theoretically define Generic Frames and assume authorial intent. We identify several routes for future research to pin down the problem: Using additional expert coding to evaluate the framing in the original articles, we find that the very implicit character of the Morality Frame resulted in problems for human coders in identifying the intended frame. We also observe that it appears elements of multiple frames show up in articles that were intended to communicate one frame, raising the question of theoretical exclusivity of frames. At the same time, we acknowledge that one study is not enough to evaluate whether it is a conceptual or methodological error - we therefore suggest that additional research is necessary to clarify where the problem is.

R2.2 The characterisation of previous work in computational frame induction is not entirely fair in tone; general claims such as "With these papers claiming validity of automatic inductive methods for detecting frames" (p.8) do some disservice to the careful discussion by essentially all the good recent computational frame induction literature as to the inherent difficulties of identifying communicative intention in text and the careful caveating of where and when such methods might be valid. Walter & Ophir (2019) invests five careful pages in this

debate, for example. Absolutely granted that analysts often ignore these caveats, but let the blame fall where it is due.

A2.2 We removed that sentence.

R2.3 I think there is a little conceptual confusion in the paper around the tasks the respective machines and human coders have been set, which could use clarification. The automated methods have been tasked with frame *induction*: no indication of what frames are appropriate has been given them, and the expectation is that they generate 5 coherent frames that map 1:1 with the 5 generic frames coded (though the authors' elegantly handle the mapping problem by assuming best case). The human coders have been tasked with frame *identification*: following the Semetko and Valkenburg instructions they are guided to identify 5 specific generic frames. If those generic frames are, in fact, present (as they are here by design) then they have been given a different and easier problem: there is no possibility of identifying *other* potential frames and thereby creating groups of text which are wrong for that reason. I can't think of a coherent way in which this can be avoided with this kind of design, but I do think it should be acknowledged.

A2.3 We added a footnote acknowledging the problem: Since human coders can work deductively, they only look for the frames they have been told to look for, while automated methods may pick up other patterns that are also meaningful, but not the frames we are looking for. We end the footnote with "For this paper, we confine ourselves to the more limited question to answer "Do our methods find the same Frames that the authors consciously put into the text?" and will bracket the more sophisticated question "If inductive methods find something else, is that something else also meaningful?"" which hopefully captures the limitation.

R2.4 The challenges created by the cost of generating this counterfactual dataset are well expressed by the authors. Formally speaking, the paper can only test the effectiveness of the methods on the particular generic frames and topics tested. Practically, the finding is probably fatal for the effectiveness of the automated methods on generic frames in general, but the strong implication of the surrounding text is that the rot goes deeper - which is difficult to determine with these data. We know that the topic modelling methods in particular can produce more frame-adjacent categories when used on narrower-scope source datasets with less intrinsic topic variability, where the variability-extracting clustering and/or dimension-reducing methods have limited topic variation to find and so draw on the variation in the ways the single

topic is discussed as a remaining source of textual variation to extract. The authors' discussion of Guo et al. (2022) on pp.8-9 is adjacent to this. It can be both true that all methods tested fail at the kind of generic frames represented by the counterfactual dataset but have at least some utility at the "more topic-like frames" such as the safety of nuclear power plants mentioned. And some authors (not exclusively those who are creators of computational frame analysis methods!) *do want* to examine frames of this kind. The points made on pp.9-10 on how topics and frames could end up conflated in a narrow dataset are good ones and I am already wondering if I could (and could afford to) replicate this paper's method to assess issue-specific frames on the kind of narrow dataset on something like the art or refugee coverage from DiMaggio et al. or Greussing & Boomgaarden. But a slightly tighter focus on generic frames in the text would be welcome.

A2.4 Following this advice, we now clarify the limited scope of the paper: In the chapter introducing Generic Frames, we now clarify why we chose to focus on the detection Generic Frames, giving two reasons for why it is advantageous for our purposes of knowing the ground truth beforehand, and for being able to produce combinations with topics in which topic and frame are independent (a clarification responding to R1.2), and also state that this limits our analyses to the detection of generic frames. We adjusted the language throughout the paper to clarify that limited scope.

R2.5 p.12: The bottom para implies that the cited authors of the various frame detection methods would claim that their methods could detect frames *under the counterfactual conditions the authors have set up*. Many of the authors expressly reported difficulties with generic frames.

A2.5 Thank you for pointing this out. We changed the manuscript to more clearly state that we are using Generic Frames as a test case, and emphasize that in previous studies, the methods are reported to struggle with identifying Generic Frames (pp. 9-10). We now highlight throughout the paper that the focus is on Generic Frame detection, not Frame detection in general.

In the discussion (p. 26), we now emphasize that our synthetic dataset represents a particularly tough case for automated methods (and, as we have found, for manual coding as well). We suggest in the discussion to apply a similar approach to issue specific frames to assess if the limitations we find only apply to generic frames or if similar issues arise in other settings.

We removed sentences that could imply the authors of the original papers would claim their methods could detect Generic Frames under the conditions of our synthetic dataset.

R2.6 p.12: I would be interested in a k=large, perhaps k=25, analysis of the different methods, perhaps in the appendix. Given that k is rarely known a priori it's common to test wide ranges of values of k and my suspicion is that some (though *definitely* not all, on the basis of the k=5 analysis) of the framing might extract across several clusters attached to each topic.

A2.6 Please see A1.14

> *…The number suggested by R2, for example, is k = 25 and then maps these 25 clusters to 5 frames. The total number of non-empty partitions s(25, 5), i.e. Stirling number of the second kind is 2436684974110751 (or 2 x 10^{13} times of 120. Suppose the computational time is constant, it would take 109589041096 years; which is longer than the age of earth). We don't have the computational power to conduct this exhaustive search.*

R2.7 p.15: I understand why the null of CCR_max is 0.2, but the paper isn't clear about why a higher 'de facto' null of 0.3 is used. Given that the higher value has the result of making all methods tested look worse it's important to document this.

A2.7 We explained now how the de facto null CCR_max was derived. In footnote 3, we also provided an intuition on how to derive it from Table 1. In our shared code, we will show how this value is computationally derived from the corpus.

R2.8 p.16: I recognise that *given the methods the authors are testing* failure is, in fact, likely to be a result of detecting topics rather than frames. But is this intrinsic to the evaluation method or just an assumption? Walter & Ophir and Nicholls & Culpepper both recommend a higher k (for their datasets) and there's a suggestion that we should not expect a 1:1 mapping. So is this something of a straw man?

A2.8 Please see A1.4

> *We have attempted to simulate a large k but we decided to give up and acknowledge this as a limitation in the discussion instead.*

In the text, P.27.

R2.9 p.16: Why a 89% CI?

A2.9 Please see A1.14

> *Given both Reviewer 1 and Reviewer 2 are both interested in why the credibility level is 89%, this issue cannot be solved by switching from 89% to the so-called standard of 95% (except we plead to convention, in McElreath's words), as, following McElreath's argument all levels of confidence are naturally equally arbitrary.. Also, we are not philosophers of science and we cannot provide a good justification on why one level of credibility is more defensible than another level. What can be said, however, is that in order to underscore this arbitrary nature of confidence levels employing a 89% CI has become somewhat of a convention in Bayesian modeling.*

R2.10 p.17: Nicholls & Culpepper didn't propose STM; they tested it as part of a range of other methods previously proposed and found it ineffective on a broad dataset which was still more focused than the multi-topic one tested here. They proposed a human-assisted analysis method starting with an an STM with a much higher k as an alternative.

A2.10 We clarified that STM was not proposed by Nicholls & Colpepper. For the query about a larger k, please see A1.4.

R2.11 p.21: The multiverse approach is also not entirely kind to those scholars who have, in fact, carefully specified a pre-processing pipeline. Walter & Ophir specified stopwording, lowercasing, depunctuating and removing numbers, and document proportion filtering to 0.5% < x < 99%. Lemmatising was rejected for theoretical reasons. Alpha was not reported. Why the testing of methods using hyperparameters and pre-processing steps that the authors have expressly rejected? I do appreciate the quasi-bootstrap nature of the multiverse procedure but it shouldn't invent variation where none should exist. At a minimum the multiverse items reflecting the options compatible with the original authors' specifications should be highlighted.

R2.12 p.26: The authors are reporting the methods' "best case" scenario in the case of attaching frame labels to arbitrary output categories, but not in the selection of hyperparameters and pre-processing steps in the multiverse analysis, where claims of the form "x% of the scenarios are not statistically better than null" are reported rather than taking the most favourable of the many possible results.

R2.13 p.26: A conditional effect of all inductive models as shown in Figure 10 may be necessary to accept/reject H1, but it's doesn't feel analytically robust as a basis of making general statements, given that the choice is not between a manual analysis and a random automated one but between a

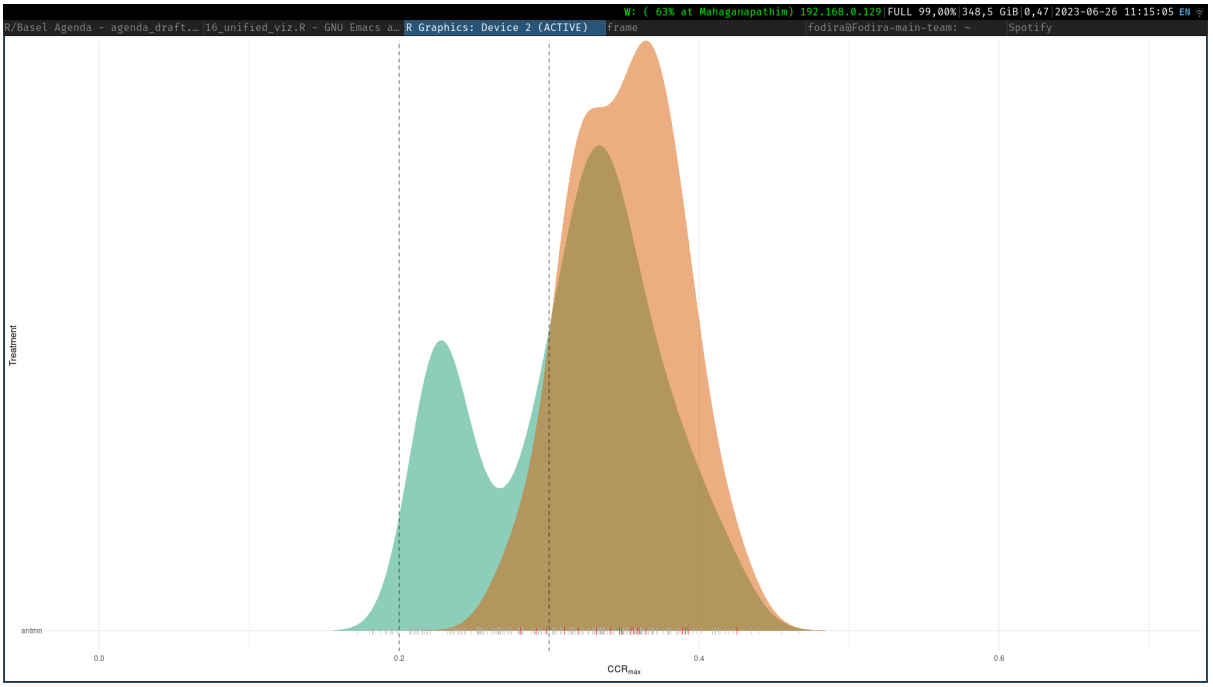
manual analysis and an automated one of the analyst's choice. A hypothetical test of 20 bogus methods and one perfect one would probably report a result crossing the null when pooled, but the takeaway would certainly not be 'the detected topics are shoehorned in as an indicator of 'frames'" in general.

A2.11-13 As R2.11-13 are about the multiverse analysis, we would like to address them jointly.

We agree with R2 (R2.11) that certain methodological decisions in the multiverse analysis are not recommended by the original developers of some frame identification methods. Also, we agree with R2 that the multiverse analysis should not be interpreted in a confirmatory manner (i.e. x% of the scenarios, R2.12) as well as in terms of central tendency (R2.13).

We remediated this by making the following changes:

1. We studied the results from the multiverse analysis in a descriptive manner (rather than using Bayesian regression) based on the density estimation in Figure 4. When it comes to studying the three hypotheses, the apogee (right tail) from one method is compared with the nadir (left tail) of another method. We believe this should be fair to all automatic methods, especially for H1. Also, we did not lump methods together into categories; we compared methods individually.
2. The central tendencies and nadirs of all automatic and semi-supervised methods are not interpreted.
3. We highlighted the recommended methodological decisions by Walter & Ophir in red in Figure 4. We did not visualize the distribution of CCR_{max} using the recommended methodological decisions because the apogees appear to be the same (Orange: Recommended; Green: Other decisions).



R2.14 p.26: As mentioned above, I think the weak finding on two different kinds of human coding is as interesting as the even weaker finding for the automated methods. It's indicative of some conceptual problems with the enterprise which could use illumination (cf Klaus Krippendorff on persistently low alpha results suggesting that the categories coded may be incoherent).

A2.14 Thank you for this comment, we agree that the issue could be a conceptual problem with how Generic Frames are described in theory. At the same time, we cannot rule out measurement error on our side, or that the dataset communicates frames less clearly than professional journalists would. Responding to this and R2.1, we now debate possible theoretical and practical problems, and what they could mean for Generic Frames as a theoretical concept while cautiously suggesting further research. Moreover, as extensively described in our response to R1.1, we have intensified our discussion of aspects that might have led to the initially observed human coding outcome, for instance by adding another analytical step in which the expert coders were confronted with and asked to judge the "ground-truth" frame for the different articles. Indeed, results of this procedure indicate that it is the Morality Frame category in particular which decreases the quality of human coding.

R2.15 p.27: It may be "turtles all the way down", but I don't think this is a serious problem for the method (as opposed to the field) here. Many texts are written by people who are not professional journalists and, given that the students were expressly asked to encode an intention, this test dataset will be as good as we could reasonably hope for for testing! If the students can't communicate their intentions, that's

simply more evidence that looking for evidence of intention in written communication is a problem.

A2.15 No action has been taken.

R2.16 p.28: The authors could presumably bootstrap CIs around their point estimates for the classifiers themselves, though in this case it would just illuminate that 100 test items is few to statistically demonstrate difference. I agree that the sample size is reasonable given the constraints and that this is simply difficult to avoid.

A2.16 Based on R1.14 and R2.9, we decided not to provide any interval estimation.

R2.17 p.28: The authors do not expressly state that a method is better or worse than null, but they strongly imply it with some of the text in the results section.

A2.17 We removed this kind of language altogether and adjusted the tone of the entire article.

R2.18 p.29: The simulation in the appendices is clever. Although it can't remove these concerns, it's a good attempt at addressing them and the authors are to be commended.

A2.18 We decided to deemphasize the benchmark of automatic / semi-supervised methods due to the unshakable concerns about small n and small k (please see A1.14).