We thank the reviewers for their helpful comments. Overall, reviewers said the work was **timely** (R2, R3) and **important** (R1, R3), with R3 adding: "the impact of the released model and released dataset is likely to be huge in the community."

We are excited that reviewers rated our Grover generator as being a **simple** (R3) yet **novel** (R1, R2) approach for multi-field document generation. Reviewers noted that Grover generates "extremely credible" articles (R2) and that due to its inductive bias as a generator (R3) it is also a state-of-the-art detector of neural fake news.

Last, reviewers said that our experiments on neural fake news detection in a semi-supervised setting were extensive (R1) and "very well done and exciting" (R3), particularly analysis about when the adversary has more resources.

Reviewer 1

(Contribution 3, Comment 1) / (Clarity Clarification 1): **machine authorship vs. fake?** Good point, we will clarify our terminology in revision. Just to clarify, Grover is pretrained only on human-written truthful news; propaganda websites are excluded. We did additional experimentation and found that Grover discriminates between human-written real news vs. human-written propaganda with **98% accuracy**, and will add this (and discussion) to the revision.

(Contribution 3, Comment 2): **evaluating consistency with metadata?** Good idea: we did additional human evaluation on the consistency of the article body with the headline, date, and author. We found that **generations are largely consistent** overall. For instance, human propaganda articles are consistent with the headline with an average score of 2.85/3 (higher is better) but machine-written propaganda gets 2.64/3; the news scores are similar.

(Contribution 3, Comment 3): **spotting generations of other models?** We ran additional experiments here: Grover detects GPT2-generated news (released by OpenAI) with **96% accuracy**, even without finetuning on GPT2 generations.

(Philosophical comment): **fake news term?** We appreciate this point and will revisit the word choice. In our draft, we followed the lead of the following political science paper, which has an extensive discussion about recommended terminology: *Lazer, David MJ, et al. "The science of fake news." Science 359.6380 (2018): 1094-1096.*

(Contribution 1, Q): **All fields share the same vocabulary** - we'll clarify this in revision. We haven't seen the model generate field-inappropriate tokens (like an invalid date), perhaps due to extensive pretraining and nucleus sampling.

Reviewer 2

Novelty of Grover over GPT2? We believe that our "novel way to guide generation" makes Grover novel, not just an 'adaptation'. Indeed, GPT(2), BERT, XLnet, and Grover share the same backbone but learn from different objectives.

What is given to the turkers? We will provide the full prompt in revision along with other details (we used 3 annotators) and discussion. For overall trustworthiness for instance, we asked "Does the article read like it comes from a trustworthy source?" Thus it emphasizes style, while content sensibility measures whether the semantic content is believable; thereby **disassociating style vs. content**. The results in the paper show that news written in a propaganda style appears less trustworthy, but the *content* of human and machine propaganda is equally sensible. The real news→Grover news results likely generalize to propaganda→Grover propaganda, were this desired.

"It takes a thief to catch a thief"? We thank the reviewer for this comment and will revise the paper to be more precise. However, we *did* test this hypothesis beyond Grover. We trained our BERT model on RealNews in a multi-field setting. Nevertheless, BERT is worse at neural fake news discrimination compared with Grover. We found this surprising because NLP leaderboards for discriminative tasks, like GLUE, show the dominance of deep bidirectional models like BERT over unidirectional ones like GPT. Even though Grover cannot handle right-to-left dependencies, it still is a state-of-the-art discriminator because its inductive bias matches that of a generator's (R3).

Choice of sources? We used only those propaganda sites whose strong political affiliations make it so they spread disinformation, and as rated by the Media Bias Chart (an updated data-driven list of news sources in terms of bias and truthfulness). This includes extreme-left websites like naturalnews.com. We will clarify this in revision.

Reviewer 3

Examples that support the analyses? We will investigate ways to visualize the discriminator in action, including color-coding predictability like (Strobelt and Gehrmann, 2019).

Experiments on human propaganda? see response to R1 (Contribution 3, Comment 1).

'counterintuitive'? Thanks for the comment. We will change the word choice accordingly. See our response to R2 "it takes a thief to catch a thief" for why we found it interesting.

Minimal overlap between models? Good point - the discriminator is initialized from generator parameters at an earlier stage of training (L182) but measuring overlap between models is challenging. Still, in our paper we consider discriminating from a larger/smaller generator (so no parameter overlap) and even in this case Grover is the best detector.

Clarity of section 6? Thanks for the comment; we will better disentangle these two key ideas in revision.