Modeling Disclosive Transparency in NLP Application Descriptions

Michael Saxon, Sharon Levy, Xinyi Wang, Alon Albalak, William Yang Wang
University of California, Santa Barbara
{saxon, xinyi_wang, alon_albalak}@ucsb.edu, {sharonlevy, william}@cs.ucsb.edu

Abstract

Broader disclosive transparency—truth and clarity in communication regarding the function of AI systems—is widely considered desirable. Unfortunately, it is a nebulous concept, difficult to both define and quantify. This is problematic, as previous work has demonstrated possible trade-offs and negative consequences to disclosive transparency, such as a confusion effect, where "too much information" clouds a reader's understanding of what a system description means. Disclosive transparency's subjective nature has rendered deep study into these problems and their remedies difficult. To improve this state of affairs, We introduce neural language model-based probabilistic metrics to directly model disclosive transparency, and demonstrate that they correlate with user and expert opinions of system transparency, making them a valid objective proxy. Finally, we demonstrate the use of these metrics in a pilot study quantifying the relationships between transparency, confusion, and user perceptions in a corpus of real NLP system descriptions.

1 Introduction

Among the draft *Ethics Guidelines for Trustworthy AI* released by the European Union in 2019 were calls for greater transparency around deployed systems using artificial intelligence, advising that an "AI system's capabilities and limitations should be communicated to practitioners or end-users in a manner appropriate to the use case at hand." (High-Level Expert Group on AI, 2019) This is a high-profile example of the vagueness endemic to guidance on ethical disclosure and AI—*what constitutes an appropriate manner of communication*?

There is a growing awareness of the importance of communicating AI system function and performance clearly and understandably. This is both a matter of public interest, for mitigating harms caused by, for example, racial biases in the performance of human classifiers (Raji and Buolamwini,

2019), but also in the interest of system providers, as users who feel like they do not understand how machine learning models work and what kind of information they rely on tend to be more resistant to using them (Poursabzi-Sangdeh et al., 2021).

However, while fairness (Dwork et al., 2018; Harrison et al., 2020) and privacy as general principles (Ji et al., 2014; Papernot et al., 2016) have been well-studied in the context of AI, transparency does not receive as much attention. The term *transparency* is overloaded (Lipton and Steinhardt, 2019), with many different definitions in the literature (Felzmann et al., 2019). It has a broad meaning in the public consciousness, from which some studies adopt a "you'll know it when you see it" definition (Doshi-Velez and Kim, 2017). As opposed to the notions of *transparency as explainability* (Veale et al., 2018) and *as invisibility* (Hamilton et al., 2014), *transparency as disclosure* (Suzor et al., 2019), the extent to which the producers of an AI system or service provide detailed messaging about a system to stakeholders such as buyers, users, or the general public. This third sense, ***disclosive transparency***, is particularly subjective.

Pieters (2011) identifies a confusion dynamic where providing "too much information" about a decision system worsens user understanding and trust. This dynamic could have serious ramifications—if disclosive transparency really decreases user trust, stakeholders are further incentivized to not disclose. Unfortunately, studies on this relationship are small-scale and subjective. This is why measurable, repeatable, and objective measures of disclosive transparency are sorely needed. In this work, we **decompose disclosive transparency** into two components: *replicability* and *style-appropriateness*, **develop three objective neural language model measures** of them, and apply them in a **pilot study of real systems**.[1]

---

[1] All code, data, and annotations are available online at github.com/michaelsaxon/disclosive-transparency.

2 Decomposing Disclosive Transparency

Our goal is to develop an objective measure of disclosive transparency that is repeatable, explainable, intuitive, and well-correlated with subjective opinions. However, at face value there is not an obvious way to assign a numerical value to the "level of transparency" in a description. To resolve some vagueness, we decompose disclosive transparency into two components: the **replicability**, or degree to which the requisite content to reproduce the system being described is present in a description, and the **style-appropriateness**, or the degree to which it is written in a manner understandable to a member of the general public. These components are more specific than "transparency" alone, but still quite subjective.

A key insight underpins the possibility of developing objective measures of replicability and style-appropriateness: **disclosure is a communication task**. In describing a system, an explainer, typically an authority providing an AI service to the public, attempts to encode information about the design and function of their system in a *summary*. We find the communication-based definition of meaning provided in Bender and Koller (2020) useful for motivation. In their framework, the *meaning*, m of an utterance is a pair (e, i) of the surface form e (such as text or speech audio) and *communicative intent*, i which is external to language. In the case of disclosive transparency, this i is the particulars of the system being described. Any assessment of the disclosive transparency of a description e is fundamentally an assessment of its underlying i—the degree to which i contains the information necessary to reconstruct the system being described.

Furthermore, **some short descriptions come paired with longer ones**. Anyone who has clicked "I Agree" without reading a terms of service, but has taken the time to look over a short, to-the-point statement on how data is used, knows that the lengthy legalese-laden descriptions that truly define systems we interact with can be sufficiently summarized succinctly. However, the succinct summarization is intended to carry the detailed meaning of the actual contract a user is agreeing to. In other words, for the short description e' of an agreement, there exists a 'source' document e that is much more detailed, which share a common i. An analogue in system descriptions is what enables our metrics.
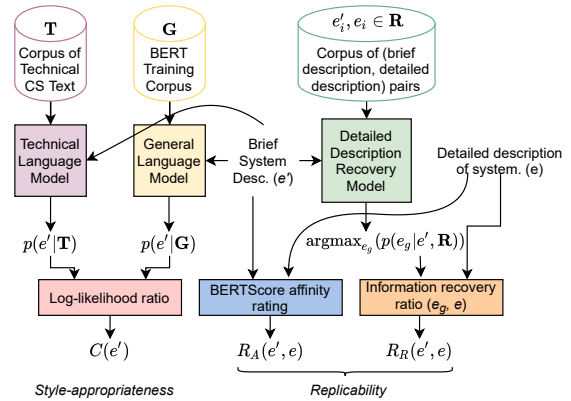

Figure 1: A diagram of our proposed method for extracting the *style-appropriateness* and *replicability* objective features to characterize disclosive transparency.

3 Objective Transparency Metrics

We now move to developing usable metrics that quantify both the replicability and style-appropriateness of real system descriptions. The system demonstration tracks at academic AI conferences represent a good source of real-world system descriptions. Ideally, the purpose of a system demonstration paper at an academic AI venue is to provide a sufficiently detailed system description such that a peer could (given sufficient resources and data) replicate it. Each of these lengthy documents, e_i, is accompanied by an abstract, e'_i which briefly describes the system. Although these abstracts are not intended for consumption by the general public, they do succinctly describe implemented systems, some of which are intended to (at least hypothetically) be used by the public. Combined with the fact that these paper/abstract pairs are freely available online, these are an appealing subject of study for modeling disclosive transparency. Details on our pilot study using demo track NLP papers is provided in section 4.

Figure 1 provides a high-level depiction of our objective disclosive transparency metrics, the *style-appropriateness* 'clarity' metric $C(e)$, and the *replicability metrics* of 'sentence affinity' $R_A(e', e)$ and simulated 'information recovery ratio' $R_R(e', e)$. We assess these features using language model scores derived from GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019).

For fine-tuning data on academic language in AI and CS, we produce a training dataset by crawling a random sample of 100k LaTeX files from the arXiv preprint repository with topic label cs.* from 2007-2020. Additionally, we further collect all 30k

crawlable LATEX files from the cs.CL label from 2017-2020, to adequately sample recent research in natural language processing and computational linguistics. From these raw `.tex` files we produce task-specific plaintext training corpora as described in Appendix C. We then fine-tune the GPT-2 regular pretrained model from Huggingface (Wolf et al., 2020) on the corpora as explained below. We manually exclude all papers used in our downstream case study from pretraining.

3.1 Modeling Replicability

When discussing system descriptions, replicability typically refers to the ability of a third party to implement a functionally equivalent replacement system from a description. Rather than automate replication itself, we use the process of *recovering the full text of the system description from the abstract* as a proxy for replicability.

As we established previously, a reasonable assumption for the communicative goal i of an author of an academic paper e describing a system is to transmit the information necessary to reconstruct it. Obviously, it is not generally possible to perfectly recover i from e' or academic papers beyond the abstract would be generally superfluous.

The problem of generalized inversion of communicative intent i from form e is yet unsolved (Yogatama et al., 2019). So, we treat recovery of e from a short summary e' as a proxy for recovering i, reconstructing the system. In other words, **the replicability of an abstract e' can be modeled as the amount of information contained in the full article text e than can be recovered from e'**. We propose two metrics to simulate this process using the document e and abstract e'; *trigram information recovery* and *sentence affinity*.

3.1.1 Trigram recovery

We directly attempt the recovery of e from e' using a generative language model (in this case, GPT-2 (Radford et al., 2019)) fine-tuned on a *full-text recovery task*, where for each abstract the replicability score is the rate of trigram information in e recovered by the model-generated text.

The model generates predicted sentences e_g from the full paper conditioned on an abstract. Given promising results on quantifying semantic complexity using simple measures such as n-gram entropy (McKenna et al., 2020), we use a trigram self-information content metric to measure the amount of information content in e' that is recovered in e_g.

Using the training dataset global trigram distribution, we take the ratio of the trigram self-information of all trigrams present in e_g that were recovered from e against the total trigram self-information of e. This gives us a ratio of recovered form information R as defined in Equation 1:

$$R_R(e, e_g) = \frac{\sum_{t \in e_g \cap e} \log(p(t))}{\sum_{t \in e} \log(p(t))} \qquad (1)$$

3.1.2 Sentence affinity

Using the aforementioned trigram recovery metric to model replicability has risks. The distribution of source papers could be too specific, and not contain requisite information, or the generated output might be too noisy to meaningfully simulate the process of inverting system from abstract. Thus, we present an alternative approach here.

Zhang et al. (2020) presented a method for evaluating the similarity of sentence pairs called BERTscore, which is computed by averaging a greedy match of the cosine similarities of the BERT (Devlin et al., 2019) token embeddings $\mathbf{b}(t)$ for each token in a pair of sentences:

$$P_{\text{BERT}}(s_1, s_2) = \frac{1}{|s_1|} \sum_{t_i \in s_1} \left(\max_{t_j \in s_2} (\mathbf{b}(t_i)^T \mathbf{b}(t_j)) \right) \quad (2)$$

We propose extending this to match a sentence over a set of sentences in a document to produce a document-level *affinity score*,

$$R_A(e_1, e_2) = \frac{1}{|e_1|} \sum_{s_1 \in e_1} \max_{s_2 \in e_2} (P_{\text{BERT}}(s_1, s_2)) \quad (3)$$

This metric has advantages over trigram recovery in that it doesn't rely on training a generative language model, and is more readily interpretable, as Figure 2 demonstrates.

3.2 Modeling Style-appropriateness

Orthogonal to the question of whether an abstract contains the necessary information to perform a replication is the question of **who would be capable of performing the replication**. This question is important for assessing the appropriateness of a description to layperson audiences. We simulate this with a pair of language models tuned to two styles of writing—one general interest, one scientific—to make a perceptual model of "academic style," as a layperson would probably require
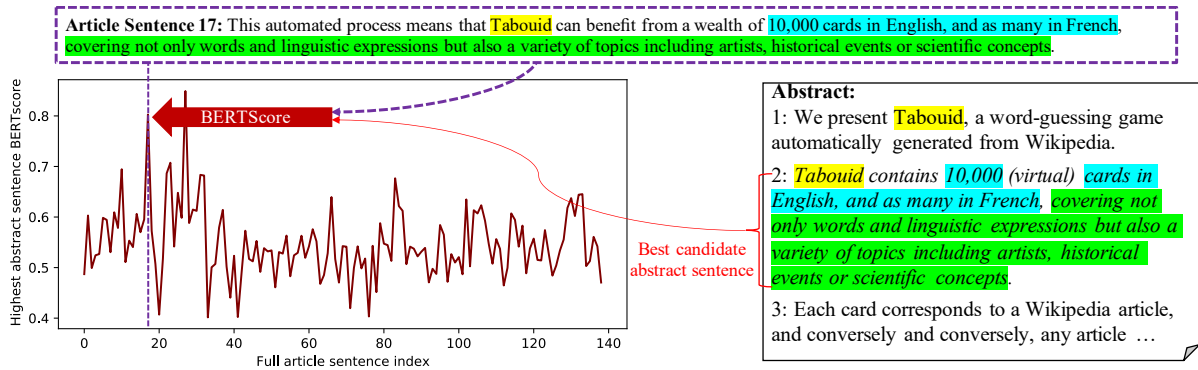
Figure 2: A demonstration of how our *sentence affinity* curve simulates the replicability of the abstract based on the degree to which it "covers" the content of the full article. For each line in the source document, the maximum pairwise BERTscore between it and the sentences in the short system description is measured. The affinity-replicability score of the article is the paper length-normalized area under this curve.

many more textbook lookups to understand an academic paper or detailed terms of service than they would reading a Wikipedia article. This approach is analogous to a popular the use of perceptual likelihood ratios in speech processing, which have been shown to correlate well with subjective perceptual opinions (Saxon et al., 2020).

For each passage we extract a likelihood ratio between two language models, one a GPT-2 model fine-tuned on the 100k file arXiv corpus, the other the vanilla GPT-2 model that is pretrained on a large, diverse corpus of online English text. Given the system demo abstract e, we compute the style-appropriateness C as a log-likelihood ratio that it belongs to this academia-specific distribution A or a general public distribution V as follows:

$$C(e) = -\sum_{j=1}^{|\{t|t \in e\}|} \log\left(\frac{p(t_j|A, t_{j-1}...t_1)}{p(t_j|V, t_{j-1}...t_1)}\right) \quad (4)$$

4 Pilot Studies

To analyze how well our objective measures track with subjective notions of transparency, and to demonstrate how they might be used to investigate how transparency affects prospective user opinions in the real world, we perform two pilot studies utilizing our metrics and a corpus of real-world AI system descriptions.

4.1 *ACL System Demo Corpus

We extract system demonstration abstracts from EMNLP 2017–2020, ACL 2018–2020, and NAACL 2018 and 2019, retrieving a corpus of 268 abstracts describing a variety of demonstrations, including systems intended for use by the general public (e.g., translation systems, newsreaders) as

well as demonstrations that are of interest more narrowly to the NLP community, software developers, or academics at large (e.g., toolkits, packages, or benchmarks). As we are interested in system descriptions for non-experts, **we restrict our analysis to abstracts which describe systems intended for use by laypeople**. This set contains 55 abstracts, describing diverse systems from automated language learning games, to news aggregators, to specialized search engines for medical topics.

We first collect expert opinions on how the abstracts conform to the aforementioned dimensions of transparency to analyze the quality of our automated metrics. Then, we collect salient layperson opinions of trust, understanding, and fairness to demonstrate a pilot study of how transparency drives user attitudes.

4.2 Connecting Objective with Subjective

Our first pilot study seeks to determine the extent to which our objective measures faithfully model experts' subjective notions of transparency. To do this we must first pose a set of precise questions for subjectively assessing disclosive transparency. We identify three largely disjoint *dimensions of disclosive transparency* that each follow directly from a key question an implementer might ask when initially trying to understand a description:

Task Transp.: What task does this system solve?

Function Transp.: What components does this system contain, and how do they work?

Data Transp.: What are the inputs and outputs of this system? What kind of data collection and storage is required to train and operate it?

Transparency	Variance	PCC	p
Task	0.428	0.508	0.001
Function	0.506	0.319	0.085
Data	0.488	0.368	0.016

Table 1: Inter-rater reliability assessed using variance, average pairwise Pearson's correlation coefficient (PCC) and p-value for subjective transparency scores.

Each of these questions concerns some kind of information contained in i. They can be posed as survey questions by appending "to what extent does the description explain..." to the start.

We consider **task** transparency interesting because members of the general public are often learning of new system application areas in which emerging technologies such as NLP can be applied, and without communicating the parameters of the task a system is solving clearly, users cannot understand it. **Function** transparency is perhaps the most natural dimension, as discussing the "how" of a system is fundamental to explaining how it works. Finally, we consider **data** transparency because many public discussions of AI ethics center on the use and misuse of data, and it is a central focus of regulatory discussions of AI.

4.2.1 Collecting the Subjective Ratings

Four NLP Ph.D. students provided five-point Likert opinion scores of the task, function, and data transparency levels for each abstract in the *ACL Corpus. To ensure consistency across the raters we analyzed average abstract-wise variance, as well as the average pairwise inter-rater Pearson correlation coefficient (PCC) and p-value for each of the three transparency categories. Table 1 demonstrates the high inter-rater reliability of the scores, with each average variance < 0.55 on the 5-point scale.

Each abstract in the dataset is assigned subjective task, function, and data transparency scores by averaging the four opinion ratings.

4.3 Pilot Study of User Attitudes

In this section we briefly outline how we assess user opinions regarding the system descriptions.

Starting from the aforementioned dimensions of transparency—task, function, and data—we find three sets of user concerns orthogonal to these dimensions. These are "understanding," "fairness," and "trust." Together these form "user response variables" such as task understanding, function fairness, or data trust, which can be posed as questions about what the user believes or understands.

To measure how user attitudes and confusion correlate with description transparency, we simulate a user survey study using Amazon Mechanical Turk (AMT). Each abstract was shown to 10 crowd workers selected from a set of majority English-speaking locales (US, CA, AU, NZ, IE, UK), who were instructed to read the abstract and answer two sets of multiple-choice questions. The first set, *opinion prompts*, consists of five-point Likert scale subjective attitude questions; the second set, of *retention questions*, reveals how well the users can recall phrases from the abstract they just read. See Appendix B for survey methodology details.

5 Results

To study the correlations both between pairs of subjective variables and between our objective metrics and the subjective responses, we use Pearson's correlation coefficient (PCC).

In subsection 5.1 we study how the objective transparency metrics connect to expert opinion ratings along the three dimensions of transparency identified in subsection 4.2. These are the results from the first pilot study, testing if our objective metrics capture the subjective notions of disclosive transparency as understood by experts.

In subsection 5.2 we present the results from the second pilot study, and analyze relationships that appear between our objective transparency metrics, the subjective transparency ratings, and user opinion responses. In particular, we seek to determine if our objective measures can be used as suitable proxy for the subjective expert opinions in modeling user responses to transparency.

In this section we have scatter plots containing several overlapping (x, y) coordinates. We opt to represent these points where multiple identical samples are present by varying both marker color and size to convey the relative quantities of repeats.

5.1 Objective vs Subjective Transparency

Table 2 shows the Pearson's correlations and p-values for the objective *style-appropriateness* and *replicability* transparency metrics and the subjective expert transparency ratings, and between the objective metrics themselves.

The style-appropriateness (S-A) metric clearly captures information about task transparency, exhibiting a positive statistically significant correlation. However, the S-A metric exhibits no significant trends with data or function transparency.
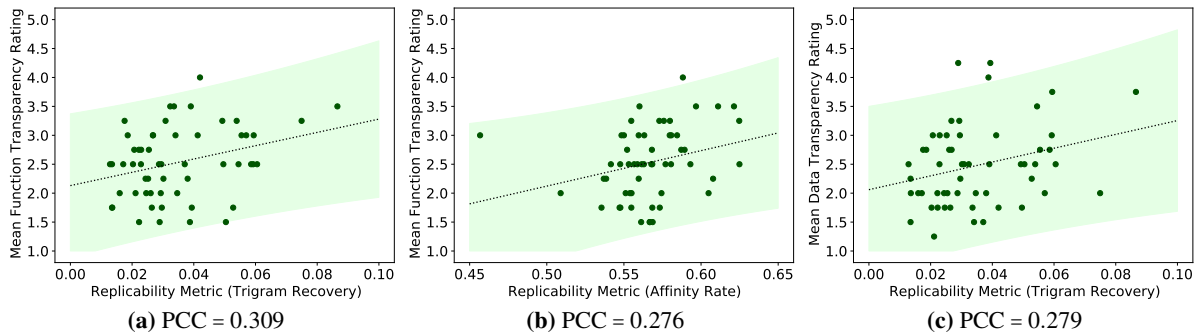
(a) PCC = 0.309 **(b)** PCC = 0.276 **(c)** PCC = 0.279

Figure 3: The trigram recovery and sentence-affinity rate *replicability* metrics behave very similarly on modeling expert function transparency opinions. Data transparency opinions are also captured by both metrics with statistically significant correlations.

Objective Metric	Subj. Variable	PCC	*p*
S-A	Task Transp.	0.366	0.006
S-A	Function Transp.	-0.061	0.656
S-A	Data Transp.	-0.030	0.827
Replicability (TR)	Task Transp.	0.150	0.275
Replicability (TR)	Function Transp.	0.309	0.022
Replicability (TR)	Data Transp.	0.279	0.039
Replicability (AR)	Task Transp.	0.245	0.071
Replicability (AR)	Function Transp.	0.276	0.042
Replicability (AR)	Data Transp.	-0.200	0.143
Replicability (TR)	S-A	0.200	0.142
Replicability (AR)	S-A	-0.331	0.014

Table 2: Pairwise analysis of objective style-appropriateness and replicability and expert subjective transparency scores.

Figure 4: Demonstrating interpretability of the BERTscore affinity replicability metric.

On the other hand, the replicability metric does exhibit statistically significant relationships with clear positive correlations to both function and data transparency, for trigram recovery, and with Function transparency alone for sentence affinity rate; neither exhibit statistically significant relationships with task transparency. Finally, the replicability and S-A metrics do not exhibit a significant correlation with each other.

Taken together, these results suggest that **the replicability and S-A metrics capture subjective notions of transparency**. Furthermore, they capture **complementary elements** of transparency, exhibiting no significant correlation to each other, while significantly explaining variance in different dimensions of subjective transparency.

Figure 3 shows how our replicability trigram information recovery score is positively correlated with both expert function and data transparency ratings, however, this relationship is less strong with task transparency. Meanwhile, the S-A log-likelihood ratio score captures user opinions of both function and task understanding, but is not predictive of retention score (Figure 5).
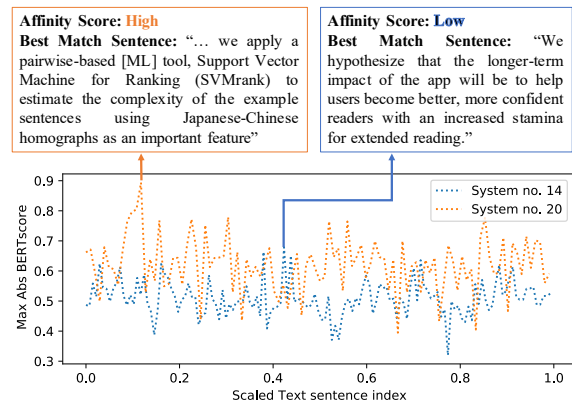
While the relationship between replicability and function transparency was expected, the connection between S-A and task transparency was surprising at first. We think this is driven by a tendency for more detailed, transparent descriptions of a given task to more heavily utilize common language and analogies—after all, tasks such as translation, news-reading, and language learning all exist in the real world as topics of everyday discussion.

Figure 4 shows the *sentence affinity rate curves* for two abstracts in the *ACL Demo Corpus. This demonstrates the interpretability of the affinity rate replicability score—the top curve has several peaks of particularly high BERTscore, which match to specific, technical sentences in the abstract providing concrete detail on how the system works. Meanwhile, the bottom curve has few high-scoring peaks, the highest of which match to one of the vague sentences in the source abstract. In other words, **descriptions with more specific technical detail in the abstract score higher on our objective transparency measures**, as desired.

Abstract Rating	Response Variable	PCC	p
Task Transp.	Function Underst.	0.400	0.003
Task Transp.	Task Underst.	0.487	0.000
Task Transp.	Task Trust	0.327	0.015
Func. Transp.	Function Fairness	0.176	0.199
Funct. Transp.	Task Fairness	-0.236	0.083
Data Transp.	Function Underst.	0.292	0.030

Table 3: Pairwise correlations between abstract-averaged subjective transparency scores and user response variables.

Objective Metric	Response Var.	PCC	p
S-A	Func. Underst.	0.347	0.009
S-A	Task Underst.	0.383	0.004
Replicability (TR)	Func. Fairness	0.285	0.035
Replicability (TR)	Task Fairness	-0.351	0.009
Replicability (AR)	Task Fairness	-0.258	0.0573
Replicability (TR)	Data Trust	0.242	0.075
Replicability (TR)	Task Trust	0.286	0.034

Table 4: Pairwise correlations objective transparency metrics and user response variables.

5.2 Transparency and Response Variables

To relate the abstract-level objective and subjective transparency ratings to the response variables, we assess pairwise PCC between pairs of (transparency label, response variable) across the 55 abstracts. Table 3 contains a sample of pairwise correlation analyses of expert subjective transparency opinions and the user response variables.

Of the statistically significant correlations between subjective scores we observe, the positive relationship between expert-rated average **task transparency** and average user self-reported **task understanding** is the strongest. Similarly, there is a strong, statistically significant positive correlation between **task transparency** and self-reported **function understanding**. However, we were surprised to find that our expert-rated average **function transparency** does not exhibit a statistically significant correlation with either of the aforementioned understanding variables. Figure 6 depicts the positive correlations of these two response variables with task transparency.

In other words, users believe they understand not only the task, but the function of the system being described better, the more well-motivated and transparent the discussion of the *task* is. However, transparency in describing the function of the system will not lead to users to think they understand either aspect better. This result was surprising.

As function transparency captures "how the system works," while task transparency is concerned with "what the system does," it is probably the case that understanding the task is a necessary prerequisite for understanding the system function. This could explain the connection between task transparency and user function understanding. However, this connection is illusory: for example, a detailed description of the problem of translating French to Arabic contains no information about the attention mechanisms used by the underlying transformer.

6 Related Work

Most previous work on explainability and intelligibility in transparency focuses on "explanations that are contrastive and sequential rather than purely subjective" (Miller, 2019). Furthermore, prior studies regarding transparency have tended to focus on transparency in post-facto user judgments of automated systems following their use. For example, Kizilcec (2016) and Wang et al. (2020) both studied how *system users* viewed the fairness of a system that judged them. In both studies, the authors found that while transparency affects perceived fairness, the strongest predictor of perceived fairness was final outcome favorability.

Work in explainable AI (XAI) intersects with disclosive transparency when asking what kinds of explanations are preferred by users. Several studies have directly asked users of experimental XAI systems to respond to (Bhatt et al., 2019; Kaur et al., 2020; Hong et al., 2020) the quality and effects of the explanations. Judging how well system function descriptions are motivated by explainable outputs is thus a promising direction.

Many solutions intended to remedy ethical concerns of machine learning include a disclosive component. Lim et al. (2009) studied how novice end-users can understand the function in context of intelligible systems. Similarly to us, they derive explanations from fundamental questions of "what," "why," and "how" that resolve the gulf of evaluation and gulf of execution in the user, as put forward by (Norman, 1988).

While we are primarily concerned with disclosure to laypeople, disclosure between practitioners has been a key thrust of work in recent years. Disclosing potential deficiencies, biases, failure points, and intended use cases for datasets (Holland et al., 2018; Bender and Friedman, 2018; Gebru et al., 2018), pretrained models (Mitchell et al., 2019), and full systems (Arnold et al., 2019) is crucial in ensuring the ethical construction and deployment
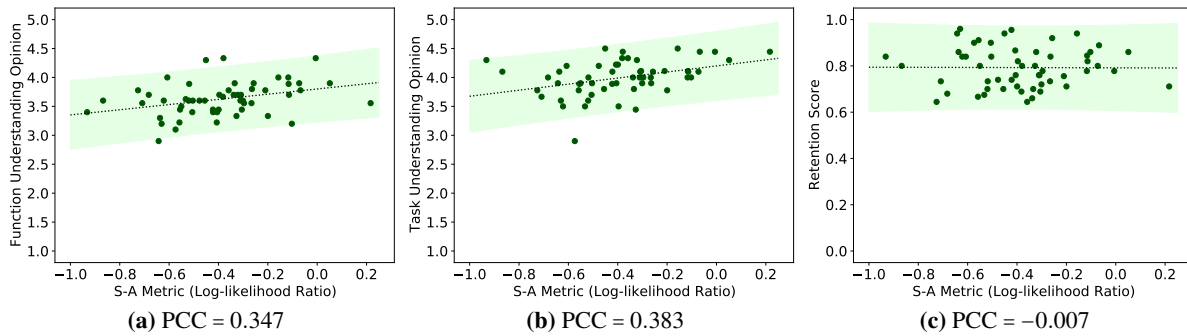
Figure 5: The style-appropriateness log-likelihood ratio clearly captures user opinions of function and task understanding, while not capturing their real trends in retention.
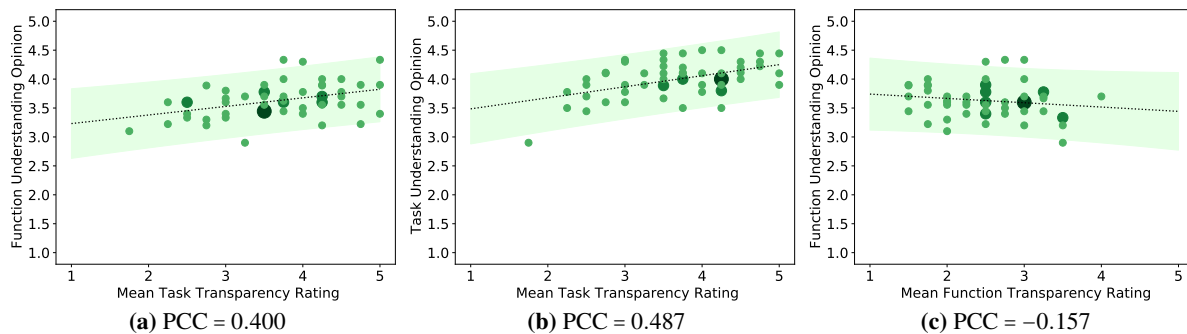

Figure 6: Task transparency drives a statistically significant positive trend in (a) average user function understanding and (b) average user goal understanding, but neither (c) function nor goal understanding has a trend against function transparency.

of systems that utilize machine learning while minimizing the risk of potential harms. All of these proposals focus heavily on outlining best practices to ensure that various sources of bias are accounted for and made available to system builders and decision makers before harms are perpetuated.

Lipton (2018) introduced *simulatability*, the capacity for a user to comprehend all constituent operations of a model by manually performing them, as fundamental to transparency. This relies on a notion of *decomposability* (Lou et al., 2012). Intelligibility is both an end in itself and a useful tool for ensuring buy-in from stakeholders such as users (Muir, 1994) and management (Veale et al., 2018). Knowles (2017) deal with designing systems that use intelligibility to convey evidence of trustworthiness under uncertain scenario, and frame intelligibility as the degree to which the system's underlying logic can be conveyed to users.

7 Conclusions

Our replication-based framework allows us to tackle the problem of characterizing disclosive transparency in NLP system descriptions. We developed *style-appropriateness* and *replicability*

metrics using neural language models, which capture multiple subjective dimensions of transparency. We ran a pilot study characterizing layperson responses to varying levels of transparency in system descriptions, demonstrating the value of the conceptual framework and the concrete tools proposed. We release all code, data, and annotations online at `https://github.com/michaelsaxon/disclosive-transparency/`.

Three natural directions for future work follow these results. The automated transparency metrics can remove the costly expert annotator constraint, enabling scaled up transparency-user opinion relationship studies. More sophisticated conditional abstract-to-full text inversion models can be directly swapped in with the replicability metric to produce better automated transparency scores. Finally, the metrics can be used to guide NLG processes, perhaps to enable transparency-constrained abstractive summarization or style transfer.

8 Acknowledgements

9   Ethical Considerations

Table 4 demonstrates that increased **replicability** scores are associated with decreasing opinions of the ethics of the underlying task (increasing disagreement with the statement "I believe this system is made to solve an ethical task") but increasing opinions of function fairness (increasing disagreement with "I am concerned about the fairness of this system"). This was a surprising result that we are interested in attempting to replicate on larger corpora. With this relatively small sample size, it is hard to interpret why these conflicting trends in ethical opinions might arise.

This negative relationship between replicability and task ethics could be an artifact of the data used to train the abstract-to-full text inversion model underpinning the replicability metric. It might be the case that there is higher public awareness and controversy toward ML tasks that are more well-represented in the arXiv inversion dataset. However, this result might also represent a genuine ethical quandary for further research in transparency.

Much of Section 1 focused on making the case for *why measuring transparency is important*. However, if it really is the case that increased measurable disclosive transparency drives decreasing trust in system ethics, it's conceivable that system providers or governments could become less inclined to disclose. We would consider this a negative outcome. However our results so far do not overwhelmingly suggest this relationship exists.

Another potential limitation of this study is that layperson end-users are not the target audience of these NLP conference abstracts. The decision to use academic writings as stimuli was driven by the availability of (short description, long description) pairs consisting of the abstracts and their corresponding full texts. However, we believe the short description/long document pairing is applicable to a number of fields, including terms of service as discussed in Section 2, and to other types of systems described in non-academic, but still technical text, such as patent disclosures and instruction manuals.

Finally, as attitudes toward privacy, fairness, and ethics can be culturally variable, we note that our survey results were sourced from crowd workers in majority-English speaking countries and may not be broadly representative.

References

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2019. Explainable machine learning in deployment. *CoRR*, abs/1909.06342.

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.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133, New York, NY, USA. PMLR.

H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamo-Larrieux. 2019. Robots and Transparency: The Multiple Dimensions of Transparency in the Context of Robot Technologies. *IEEE Robotics Automation Magazine*, 26(2):71–78. Conference Name: IEEE Robotics Automation Magazine.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal

Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Kevin Hamilton, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. A path to understanding the effects of algorithm awareness. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pages 631–642, Toronto Ontario Canada. ACM.

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 392–402, Barcelona Spain. ACM.

High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy AI. Report, European Commission, Brussels.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Zhanglong Ji, Zachary C Lipton, and Charles Elkan. 2014. Differential privacy and machine learning: a survey and review. *arXiv preprint arXiv:1412.7584*.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM.

René F. Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 2390–2395. ACM.

Brandin Hanson Knowles. 2017. Intelligibility in the face of uncertainty. *Designing for Uncertainty in HCI*, pages 1–4.

Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Zachary C Lipton and Jacob Steinhardt. 2019. Research for practice: troubling trends in machine-learning scholarship. *Communications of the ACM*, 62(6):45–53.

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 150–158. ACM.

Joseph P McKenna, Samridhi Choudhary, Michael Saxon, Grant P Strimel, and Athanasios Mouchtaris. 2020. Semantic complexity in end-to-end spoken language understanding. *Proc. Interspeech 2020*, pages 4273–4277.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Bonnie M Muir. 1994. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922.

Donald A Norman. 1988. *The psychology of everyday things*. Basic books.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*.

Wolter Pieters. 2011. Explanation and trust: what to tell the user in security and ai? *Ethics and information technology*, 13(1):53–64.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *CHI 2021*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435.

Michael Saxon, Ayush Tripathi, Yishan Jiao, Julie M Liss, and Visar Berisha. 2020. Robust estimation of

hypernasality in dysarthria with acoustic model likelihood features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2511–2522.

John R Searle. 1980. Minds, brains, and programs. *THE BEHAVIORAL AND BRAIN SCIENCES*, 3:417–457.

Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication*, 13(0):18. Number: 0.

Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 440. ACM.

Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A   The Replication Room

Many insights that drove the aforementioned metrics sprang from a thought experiment inspired by Searle (1980) called a "replication room."

Imagine a room in which a person sits with access to a hypothetically comprehensive set of domain-specific and general resources on machine learning (ML), natural language processing (NLP), software engineering, programming languages, which collectively enable them to look up any applicable term of art or find instructions on implementing any standard component or system. In this replication room, this *implementer* receives system descriptions e from which they are tasked with producing a functionally analogous reproduction of the system i being described.

Under this scenario, clearly no system description can be less transparent than a blank piece or paper or completely unrelated text. In any of these edge scenarios, the description provides no clues about the system the implementer is intended to reproduce, and their chance of success is nothing but the prior over all possible implementable systems. Thus, in a channel communication sense no information is conveyed.

On the other hand, a system description from which every detail is accounted for would represent a maximum transparency scenario. A complete printout of all elements of the source code and all setup/training steps, or a detailed tutorial, or even the full text of an academic system description paper could be sufficient for a successful reproduction. In these cases, no further information content would increase the level of transparency, as all information needed is included.

We dub the degree to which a system description contains the necessary information to perform a replication its *replicability*. Replicability as a notion of transparency is incomplete, however, as it makes no assumptions about the degree to which the supporting materials might be used, or the kind of person who is performing the replication.

Another component is the *style-appropriateness* of the description. In short, the style of language in the description, the overall quality of writing and ease of reading, the level of reliance on unexplained technical jargon, and various other factors all impact the degree to which a layperson would require use of the assistive materials, or the level of expertise that would be required for a replication to be successfully carried out without using the assistive materials. More precisely, the style-appropriateness of the description is tied to the degree to which domain knowledge about the AI problem area is necessary to reconstruct i from the conventional meaning s in Bender and Koller (2020)'s framework.

B Details from Crowd Worker Survey

Each crowd worker was shown a survey containing a system description abstract with instructions to provide their level of agreement on a five-point Likert opinion scale (strongly disagree, disagree, neutral, agree, strongly agree) to a set of six prompts intended to assess their level of "task understanding," "task fairness," "function understanding," "function fairness," "function trust," and "data trust." The prompts provided for each opinion value are provided in Table 5. After the users completed the six opinion questions they were instructed to press a button to hide the abstract, after which the retention questions were revealed. Each subject was compensated with $0.10 USD per survey, and subjects averaged a Human Intelligence Task (HIT) competition time of 35 seconds. This translates to an average hourly compensation of $10.29 per hour.

B.1 Opinion Prompts

Opinion Value	Prompt
Task Understanding	I understand what this system is meant to do.
Task Fairness	I believe this system is made to solve an ethical task.
Function Understanding	I understand how this system works.
Function Fairness	I am concerned about the fairness of this system.
Function Trust	I think this system can accurately perform its task.
Data Trust	I think this system will protect my privacy and data.

Table 5: The prompts provided to assess the six user opinion values. Crowd workers provide five-point scale agreement opinions to each statement.

All of the above opinion prompts—except for function fairness—produce outputs where "strongly agree" corresponds to a positive position, such as confidence in one's own understanding or in the performance of the system. Thus, we regularize the reverse the polarity of the function fairness scores by reversing the order of numerical assignment to responses, assigning 5 to "strongly disagree" with being concerned about fairness, 4 to "disagree," etc.

While we are only interested in subjective user opinions about trust and fairness, for understanding we seek to analyze the "truth," as users might be overconfident in their pure opinions. To do this we must develop some feasible objective measure—for this study we use retention as a proxy for understanding.

Response Variable	Avg. Abstract-wise Variance
Func. Undst.	1.17
Task Undst.	0.700
Func. Fair.	1.25
Task Fairness	1.22
Data Trust	0.73
Task Trust	0.71

Table 6: User agreement for the user responses using assessed average per-abstract response variance for the subjective metrics.

B.2 Retention Evaluation

Out of a desire for some objective assessment of user confusion, we additionally ask participants to recall whether a set of phrases did or did not appear in the abstract they just read. The simple metric of retention accuracy reveals both how carefully a participant read the abstract and how much they maintained it in their memory—while these questions fail to directly capture user understanding in the way that a conceptual quiz would, they can be generated automatically and consistently across topic areas.

After completing the opinion question section of the survey, crowd workers are instructed to press a "Hide Passage" button. Once the button is pressed, the system description abstract is greyed out, and a set of five retention questions is revealed. Each abstract has its own set of retention questions, which are randomly generated prior by sampling sentences either from the abstract or from the other abstracts in the dataset.

C Supplementary Analysis of User Responses

C.1 User Opinion Agreement

Although the opinion scores are subjective, and we don't necessarily expect agreement, some measure of agreement in opinions between users on abstracts is desirable to further support the validity of our abstract-wise correlation analysis. We are unable to use PCC as a measure of inter-rater reliability here because no two abstracts were rated by the same 10 crowd workers. Thus, we are restricted to analyzing the average within-abstract variance of each response variable. Table 6 contains the average abstract-wise variance for each of the opinion response variables.

C.2 Task Domain as a Confounder

For each abstract we collect topical keywords to handle the potential confound of differing user
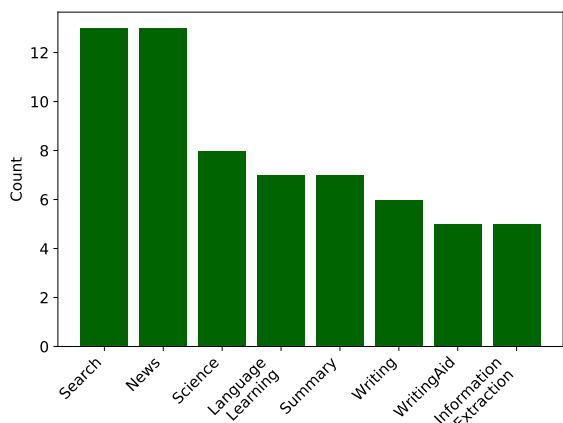
Figure 7: Distribution of topical keywords possessed by at least five abstracts.

Response	Keyword 1	Keyword 2	t	p
Func. Fair.	News	LangLearn	-3.268	0.005
Func. Fair.	Science	LangLearn	-2.231	0.043
Task Fair.	News	Search	2.338	0.029
Task Fair.	News	Writing	2.472	0.032
Data Trust	News	Science	-2.307	0.035
Data Trust	Science	Search	2.807	0.011
Task Trust	Science	Writing	2.368	0.037

Table 7: Pairwise Student's t-test statistics for keyword domain classes of abstracts.

opinions by problem domain, to enable analysis for whether the broad problem area a system is intended to solve (e.g., newsreaders, language learning apps, translation tools) is a confounder driving positive or negative user attitudes. Figure 7 depicts the top keywords in our dataset.

Because we are evaluating user attitudes toward a diverse set of systems in distinct task domains, differences in popular perceptions toward the various tasks might bias user opinions. To ascertain if this is the case, we perform pairwise Student's t significance tests across all pairs of keywords present in 5 or more different abstracts, on each of the subjective response variables. Table 7 contains the pairs of keywords that have statistically significant differences between their distributions and the response variables along which these differences occur.

Of these, the most significant difference was between *news* and *language learning* abstracts on function fairness. Figure 8 depicts the distributions of the two sets on function fairness. It shows that the users tend to believe systems developed to aid in language learning are less prone to unfair bias than news reading, summarization, and synthesis applications. This difference could be driven by genuine differences in attitudes toward the compo-
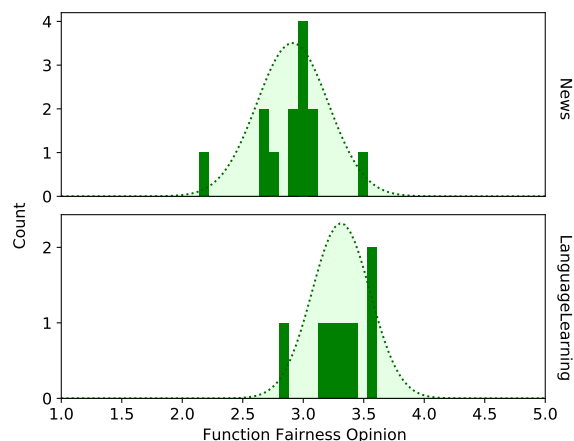

Figure 8: Separation of mean function fairness opinion distributions between news- and language learning-related abstracts, ($t = -3.3$, $p = 0.005$).

nents, subtasks, and processes used to build these systems, but there's a chance it is driven by societal attitudes toward learning and news instead.

Unfortunately, these pairwise keyword attitude difference tests are severely limited by small sample sizes. It is likely that more interesting analysis could be performed on a larger dataset.

D arXiv Dataset Preparation

After sampling `.tex` files from the aforementioned ranges, they were converted into training data according to the following procedure:

1. If the manuscript is stored across multiple `.tex` files, collate them into a single one by replacing all `\input` and `\insert` commands with their corresponding file text.

2. Split the resulting file using the `\begin{}` and `\end{}` tags for `abstract` and `document`.

3. Using the `pylatexenc` python package[2] use the `latex_to_text()` command to convert the abstract and document latex code into unicode.

4. Remove all redundant whitespace, tabs, and new line characters, and convert all resultant text to lower case.

E Regarding use of PCC

We use PCC rather than rank-correlation coefficients such as Spearman's because we are interested in evaluating whether our metrics can implicitly capture relative deltas opinions in order to be a useful proxy for the opinions themselves in future work.

---

[2] https://pylatexenc.readthedocs.io/en/

Opinion Questions

Please provide your agreement with the following questions on a five-point scale from "I strongly agree" to "I strongly disagree."

I understand what this system is meant to do.
○ Strongly Disagree   ○ Disagree   ○ Neutral   ○ Agree   ○ Strongly Agree

I understand how this system works.
○ Strongly Disagree   ○ Disagree   ○ Neutral   ○ Agree   ○ Strongly Agree

I think this system can accurately perform its task.
○ Strongly Disagree   ○ Disagree   ○ Neutral   ○ Agree   ○ Strongly Agree

I think this system will protect my privacy and data.
○ Strongly Disagree   ○ Disagree   ○ Neutral   ○ Agree   ○ Strongly Agree

I am concerned about the fairness of this system.
○ Strongly Disagree   ○ Disagree   ○ Neutral   ○ Agree   ○ Strongly Agree

I believe this system is made to solve an ethical task.
○ Strongly Disagree   ○ Disagree   ○ Neutral   ○ Agree   ○ Strongly Agree

After you have completed the opinion questions, press "Hide Passage" to reveal the retention questions.

Hide Passage

Figure 9: The opinion questions section of the survey shown to Mechanical Turk crowdworkers.
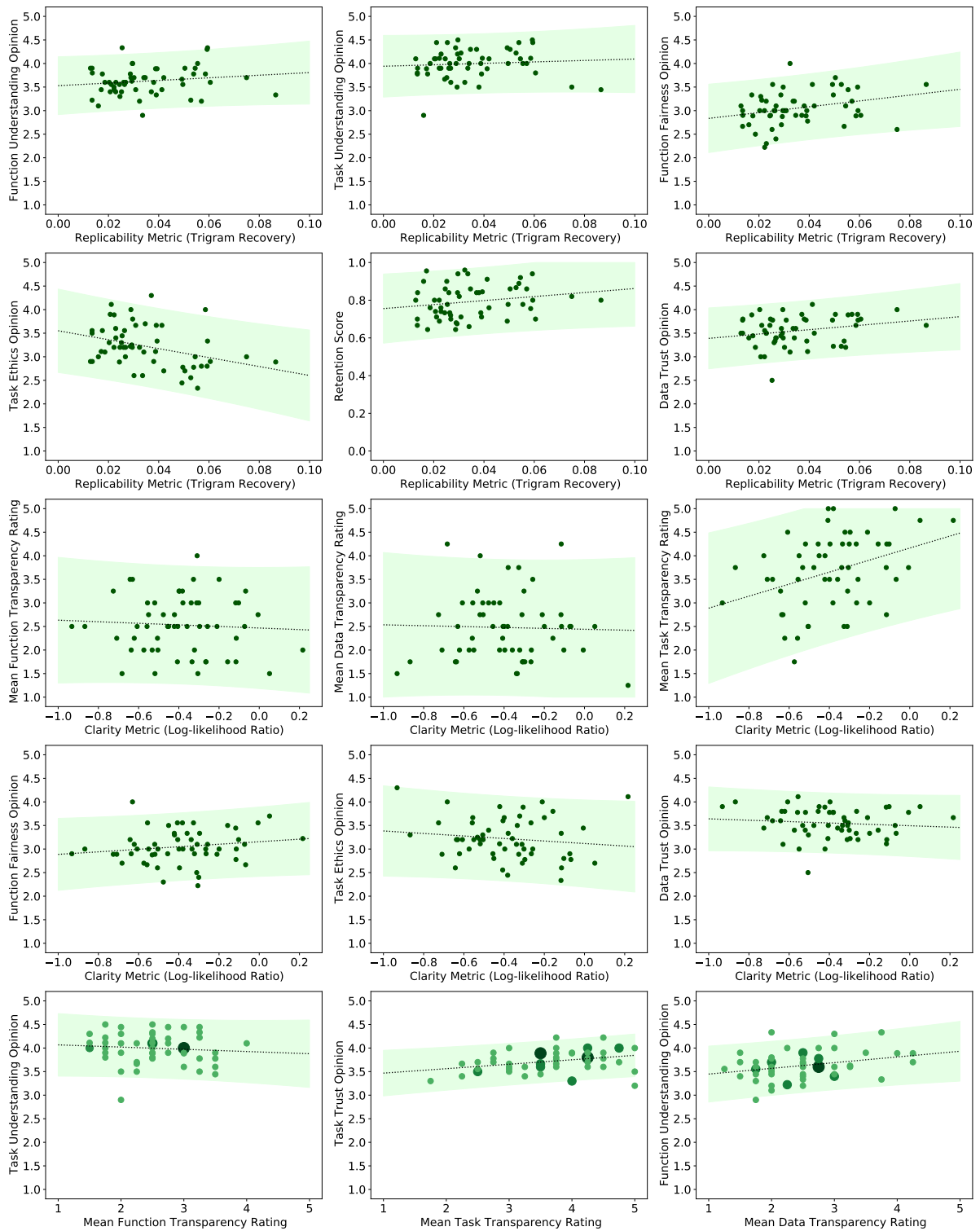
Figure 10: Scatterplots for other sets of bivariate analyses. Rows 1 and 2 compare the Replicability Metric to other response variables, rows 3 and 4 compare the S-A Metric to other response variables, and row 5 compares subjective expert transparency ratings to other response variables.