The Tournesol dataset: Which videos should be more largely recommended?

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

This paper introduces the Tournesol public dataset, which was collected as part of the online deployed platform `https://tournesol.app`. Our dataset contains a list of 1,116,318 comparative judgments of YouTube videos by 8,804 users of the Tournesol platform. 263,668 of these judgments were about which video should be more largely recommended, while the remaining evaluate secondary criteria like content reliability, topic importance and layman-friendliness. The dataset also exports information about users' pretrust statuses and vouches. It is published at `https://api.tournesol.app/exports/all` under ODC-By license. The data is currently used by Tournesol to make community-driven video content recommendations to over 6,000 users.

1 Introduction

Recommendation AIs have become extremely influential. In the last few years, beyond their impacts on mental health [58, 21, 94], because they amplify disinformation, cyberbullying and hate, they have been linked to major geopolitical events, including COVID disinformation [81, 46], the rise of far-right parties [93, 92, 98], and the Rohingya genocides [42, 73]. Crucially, in all these examples, the victims of recommendation AIs are not only their users; hate amplification is threatening entire populations, even when these populations do not use recommendation AIs themselves. This is in sharp contrast with the overwhelming majority of the scientific literature, which assumes that recommendation AIs should be optimized for their users only [1, 71].

As online activities grew, recommendation AIs have *de facto* taken the role that was traditionally played by these intermediate bodies [91, 50]. For instance, by amplifying the cyberbullying of climate scientists, Twitter's AI provoked their exodus from the platform [95], thereby turning climate change into a *mute news*, which is endangering plenty of non-users [3]. The great replacement of the intermediate body by privately owned AIs has been tied to an alarming decline of democratic norms worldwide, as many reports expose a global trend of autocratization [72, 7].

So how do today's large-scale recommendation AIs address the ethical dilemmas that they face billions of times per day, when they are tasked with amplifying some (potentially hateful) content over others (of potential public interest)? Currently, they heavily rely on (highly sophisticated) *machine learning* [26, 65]. In other words, such AIs leverage massive amounts of data to determine which content they will promote at scale. However, as an immediate corollary, such AIs are exposed to *manipulation* by *poisoning* data [89]. In fact, this poisoning has been industrialized, not only by authoritarian states [20, 48], but also by private companies based in the UK [53], Spain [16], Israel [6], France [90] and Switzerland [37]. The magnitude of this industry is well captured by one puzzling statistic: Facebook reportedly removes around *7 billion fake accounts per year* [60].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

While a recent line of research has provided numerous poisoning mitigations [15, 34, 35, 30, 83, 76], it is also known that there are fundamental impossibility theorems that prevent accurate learning in highly adversarial, heterogeneous and high-dimensional settings [31, 61, 39, 33]. In particular, *there is no substitute for training datasets of high quality and security*. In particular, to design trustworthy ethical AIs, it is essential to train them on large, secured and trustworthy datasets of human ethical judgments. In this paper, we present the *Tournesol public dataset*, whose goal is to remedy the current state of affairs. More precisely we make the following contributions.

Contributions. Our main contribution is to present and share the *Tournesol public dataset*, which can be downloaded directly from `https://api.tournesol.app/exports/all`. The dataset consists of 263,668 pairwise comparisons of the recommendability of 56,796 YouTube video by over 8,804 Tournesol accounts. Additionally, the dataset contains 852,650 pairwise comparisons of the videos' quality on secondary criteria, such as reliability, importance and layman-friendliness. Our dataset, published under ODC-By license, also contains pretrust information about contributors, vouches between contributors, as well as scores computed from the data using SOLIDAGO [14]. Crucially, the dataset was collected in a fully deployed environment with actual stakes, as Tournesol eventually makes recommendations based on the provided data to over 6,000 users.

The paper also presents an analysis of our dataset, with valuable insights for the ethics of content recommendation. One finding is that the topic importance highly matters in Tournesol's contributors' judgments. While caveats apply, this suggests that the attention to "fake news" may be misguided; in fact, the disinformation industry often proceeds *without* producing false information, e.g. by overclaiming positive impacts, shifting blame or bullying critics [77]. Prioritizing greater exposure to *mute news* might be more urgent. Our analysis also highlights the need of psychological-based preference learning models, as we expose biases and variations in contributors' judgments.

Finally, our paper discusses numerous exciting research directions that our public dataset could inspire or facilitate. In particular, we believe that a lot more focus should be given to secure learning under poisoning attacks, but also to *Proof of Personhood*, *expertise validation*, *volition learning*, *active learning* and *resilient collaborative filtering*, among others.

Literature review. Tournesol presents a new contribution to the growing field of AI alignment with human values [49, 24, 54, 79], which aims to teach human preferences to AIs, and to design systems that maximize what humans prefer to maximize [84, 56]. Clearly, this requires finding out about humans' judgments on how AIs ought to behave. Unfortunately, so far, to the best of our knowledge, all published content evaluation datasets [52, 12, 78, 100, 101, 97, 23, 8] are consumer-centric, i.e. they report what consumers prefer to consume; not what they regard as recommendable to others.

To collect such data in a realistic setting, Tournesol's dataset draws inspiration from several previous AI ethics solutions, which leveraged *collaborative governance* to address cases of conflictual human judgments. In particular, [64] introduced WeBuildAI, a framework where stakeholders of a food donation system could weigh in on the identity of the recipient of a donation. One challenge is that such decisions must be made every day; but stakeholders are not available every time a decision needs to be made. To account for their preferences, WeBuildAI asks stakeholders to either write down an AI that describes their preferences, or to provide judgments on generated food donation dilemmas. In the latter case, a learning model is then used to infer how the stakeholders would likely assess other dilemmas. In any case, an *algorithmic representative* is thereby constructed for each stakeholder; and the resulting decision will follow from a vote of the algorithmic representatives. Similar approaches were proposed for kidney donation [45] and for the "trolley dilemmas" [43] that autonomous cars could one day face [11, 75].

Perhaps most similar to our approach are Twitter's *Community Notes* [99, 80], whose governance is intended to be fully community-driven. More specifically, the system allows a community of contributors to add a note to misleading tweets, e.g. to correct misinformation or to add context to prevent confusion. The contributors cannot only propose the note; they are also asked to assess other contributors' notes. Notes that are judged helpful by a sufficiently large and diverse set of contributors are then published by the platform. The system is very transparent, and provides a lot of freely accessible data on human judgments[1].

[1]The data can be downloaded here: `https://communitynotes.twitter.com/guide/en/under-the-hood/download-data`

Structure of the paper. In the sequel, Section 2 will present our public dataset, and the context in which the data was provided. Section 3 presents an analysis of our dataset. Section 4 then provides a list of research challenges that are raised by the dataset. Finally, Section 5 concludes.

2 The dataset

In this section, we describe our main contribution, namely the release of a new, scalable, secured and trustworthy database of reliable human judgments.

2.1 Raw data

Pretrust. To guarantee the security of our data, Tournesol aims to verify that every account is owned and controlled by a human, and that this human only owns and controls this single account on the platform. In other words, Tournesol aims to obtain a *Proof of Personhood* [17] to verify each active Tournesol account, and to thereby prevent *Sybil attacks* [28]. Unfortunately, there is currently no reliable and scalable solution for *Proof of Personhood*.

Today's main solution is *email certification*. More precisely, when they create a Tournesol account, contributors are asked to validate, if possible, an email address from a trusted email domain. The list of trusted email domains is currently managed manually. An email domain will be considered trusted if it seems sufficiently unlikely that a large number of fake accounts can be created from this domain.

This excludes domains like @gmail.com and personal domains like @my-personal-website.com. The concern is not only that the domain will maliciously create a large number of fake accounts; it is also that they may be hacked by a malicious entity that will create such fake accounts. The list of trusted email domains is available at `https://tournesol.app/about/trusted_domains`. It includes domains like @epfl.ch, @who.int and @rsf.org. 795 contributors are thereby authenticated.

Evidently, however, this solution is still highly imperfect. On one hand, this does not guarantee the absence of fake accounts. On the other hand, and perhaps more importantly, this excludes most potential contributors from participating.

Vouching mechanism. To propagate trust to more accounts, Tournesol also proposes a vouching mechanism. Namely, any account can vouch for the authenticity of another account. More precisely, the account must vouch that the other account is used by a human who is not using any other account on the platform. The dataset contains 129 vouches.

Comparison-based judgments. Following a large literature on the topic [41, 19, 68, 11, 75, 64, 45], Tournesol relies on a comparison-based preference elicitation system. We believe that the need to distinguish among top content which should be more recommended makes this system more suitable than, e.g., using direct assessments [67, 2, 59, 88], which may yield too many "saturated" maximal assessments. Additionally, comparisons are labeled with the week in which the comparison was first submitted. This allows potentially observing changes or drifts in the contributors' judgments.

Figure 1 (left) presents the video comparison interface. Namely, contributors are asked to select two videos, and to tell Tournesol which one of the videos should be recommended at scale. Moreover, rather than a binary decision, the contributor is asked to provide the judgment by moving a slider on a more continuous scale, from -10 to 10, The value -10 means that the contributor would prefer Tournesol to recommend the left video vastly more often than the right videos, while the value 0 means that they believe both videos should be recommended equally often.

Quality criteria. Tournesol allows contributors to rate nine other *optional* quality criteria (Figure 1)

- **Reliable and not misleading:** Is the presented information trustworthy, robustly backed and properly nuanced?
- **Clear and pedagogical:** How efficiently does the content guide viewers in their understanding?
- **Important and actionable:** Can additional focus on this topic have a significantly positive impact on the world?
- **Layman-friendly:** How understandable is it, without prior knowledge?

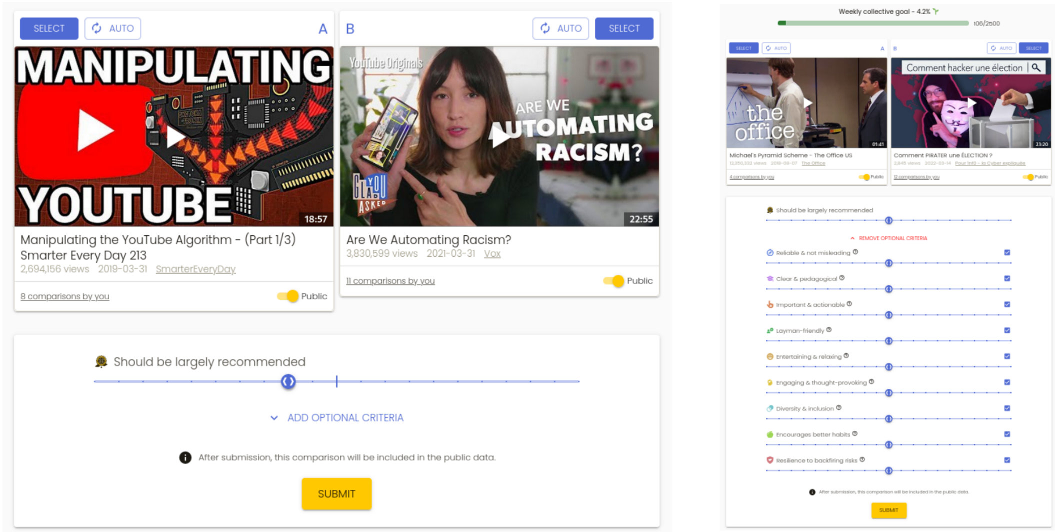


Figure 1: The interface through which contributors are asked to provide judgments. The judgments are comparisons of video contents using a slider along the main criteria "should be largely recommended" (left) and optional quality criteria (right).

- **Entertaining and relaxing:** Do people feel good watching it?
- **Engaging and thought-provoking:** Does it catch people's attention, spark curiosity and invite to question previous beliefs?
- **Diversity and inclusion:** Does it promote tolerance, compassion and wider moral considerations?
- **Encourages better habits:** Does it make people adopt habits that benefit themselves and beyond?
- **Resilience to backfiring risks:** Is it adapted to viewers with opposing beliefs? Does it prevent misconceptions or undesirable reactions?

While the criteria are further provided on Tournesol[2], most contributors have surely *not* read thoroughly our descriptions. Arguably, they will more likely judge these criteria according to their own understanding, which will be mostly based on the name of the criteria.

2.2 Processed data

In addition to the raw data presented thus far, the Tournesol public dataset exports processed data. The processing is performed by a pipeline called SOLIDAGO [14].

Solidago. The pipeline has six modules. First, pretrust and vouches are used to assign *trust scores* to all users. Second, *voting rights* are assigned to the different users, in a way that includes untrusted users, while guaranteeing that they cannot outweigh trusted users. Third, for each criterion and each user, the comparisons are turned into the user's *raw scores*, using the generalized Bradley-Terry model [36]. Fourth, raw scores are *scaled*, using Mehestan [4], zero-shift and standardization. Fifth, scaled scores are securely aggregated into *global scores*, using the Lipschitz-resilient quadratically regularized quantile [14]. Sixth, all scores are squashed into $(-100, 100)$, using the map $t \mapsto 100t/\sqrt{1+t^2}$. All along, left and right uncertainties on all variables are computed.

Exported values. Trust scores, squashed individual scores and squashed global scores are provided in the public dataset.

Results. Figure 2 lists the most recommendable videos, according to Tournesol's contributors, as they are displayed on the website.

[2] `https://tournesol.app/criteria`

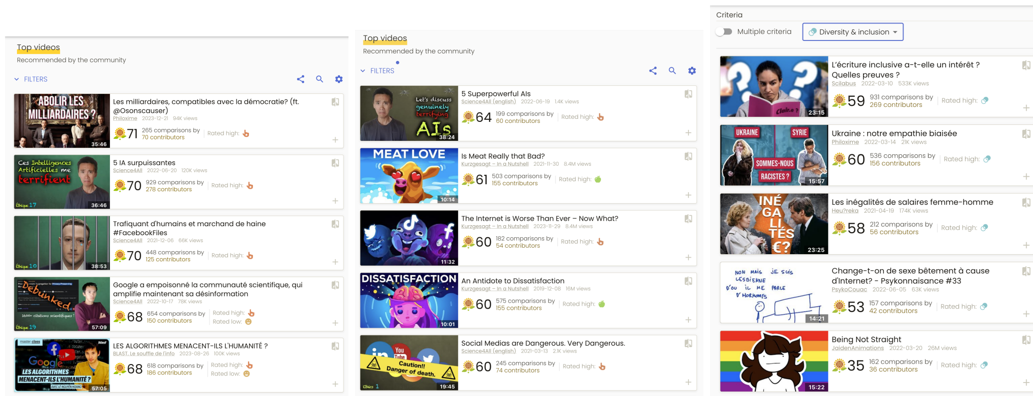


Figure 2: Best videos (left), best English-speaking videos (middle) and best videos along the criterion "diversity & inclusivity" (right).

2.3 Privacy

Overall, we encourage transparency in our contributors, as we believe that this will foster important research on human judgments, and help make safer and more ethical AIs. However, we acknowledge that, because of social and political pressures, some judgments are dangerous to make public, e.g. when criticizing one's own employer or government. This is why we allow contributors to provide data publicly or privately. More precisely, each contributor can select the privacy setting of any video they rate. If a video is rated privately, then all its comparisons to any other video will be recorded privately. Only Tournesol's server can access to such data. Conversely, all comparisons that involve two publicly rated videos are exported in the Tournesol public dataset.

2.4 Data collection context

The contributors to Tournesol receive no financial compensation. Their contributions are mostly motivated by the desire to contribute to a democratic AI governance project, and by the will to promote content of public interest. Their recruitment is thus organic, and mostly depends on how frequently they were exposed to the promotion of the Tournesol project. Evidently, this greatly correlates with Tournesol's communication, which has been heavily supported by the (French-speaking) YouTube channel Science4All, and by other science communicators [55]. As a result, the set of contributors is in no way representative of the global population. Namely, it is heavily biased towards science enthusiasts. Nevertheless, we believe that the data provided by this community should be of great interest to AI alignment, at least on topics with a significant scientific component.

3 Data analysis

This section presents some data analyses to provide insights in the *Tournesol public dataset*.

3.1 Contributors' contributions

Figure 3 displays the number of contributions per user. Perhaps unsurprisingly, this statistics is heavy-tailed; in fact, it seems to fit Zipf's law [85], with a few contributors providing most of the comparisons, and most of them providing very few. Figure 4 plots the activity through time: Tournesol has 100 to 200 weekly active users, while the number of monthly active users fluctuates between 200 and 900.

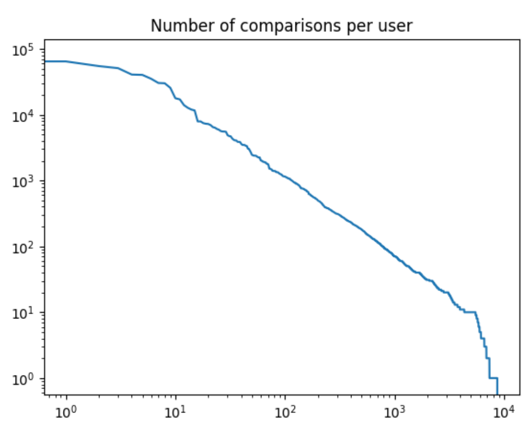


Figure 3: Number of comparisons provided by the different contributors, on a log-log scale, which is typical of Zipf's law [85].

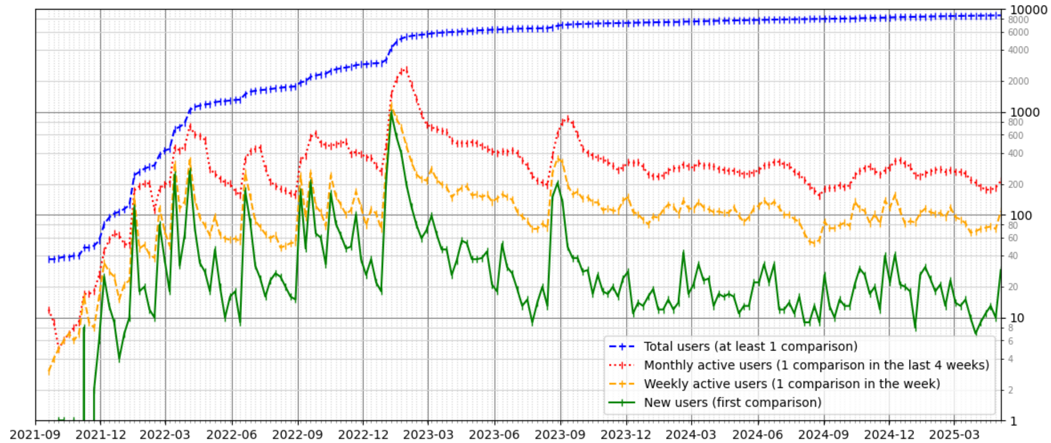
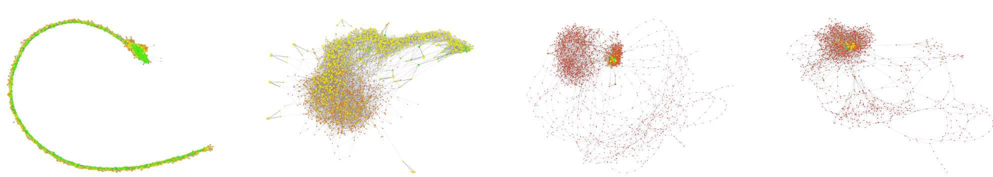


Figure 4: Contributors' participation through time.

3.2 Video and contributor connectivity

For scores to be meaningful, the contributors must have compared sufficiently many videos in common [4]. The contributor comparability graph has a connected component with 8,064 contributors and diameter of 6, out of the 8,692 contributors that have compared at least 2 videos. The graph has 256,360 edges out of 37,771,086 possible (0.68%) making it very sparse. But for the induced graph of the top 100 most active contributors with a trust at least 0.1 (which correspond to *scaling-calibration* contributors [14]), 3,699 (75%) pairs of contributors are comparable. This justifies the restriction of scaling calibration to the most active contributors.

Figure 5 details video comparisons for some highly active users. Interestingly, because the platform lets contributors to select their videos to compare, we observe a wide variety of comparison graphs. This raises open questions about the uncertainties of the resulting learned scores [36], and about the possibility to improve accuracy through *active learning* [69, 86].



(a) Contributor "scayrol" (b) Contributor "white" (c) Contributor "zekk" (d) Contributor "ThugFou"

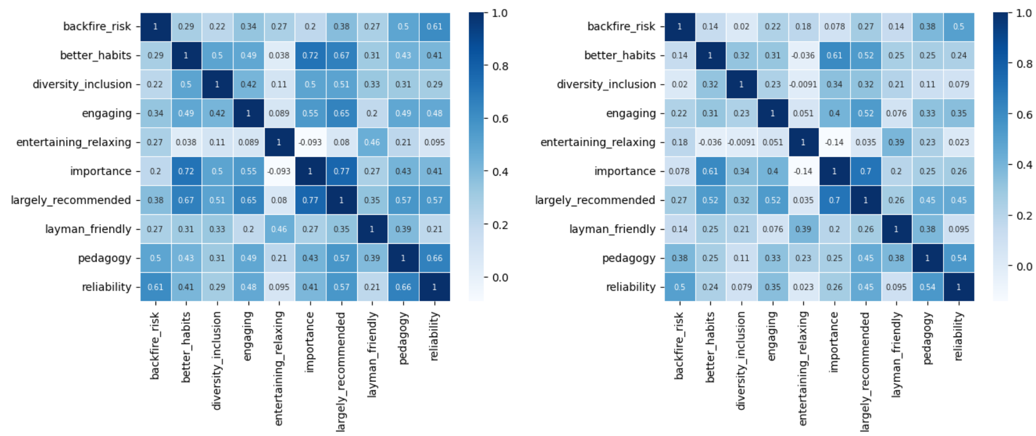
Figure 5: Graphs of video comparisons for different users

3.3 Correlations between criteria

Figure 6 reports the correlations between quality criteria, in contributors' comparative judgments. Perhaps most remarkably, we observe that the criterion that best predicts whether a video "should be more largely recommended" is whether it is "important and actionable". This finding highlights the need to pay greater attention to *information prioritization*, and especially combatting "*mute news*" [55]. In particular, there may be an excess of attention to "*fake news*". In fact, [77] expose numerous strategies from the "merchants of doubts" that do not involve producing false information, such as shifting blame, cyberbullying critics or "striking a positive tone" [27].

Figure 6 also shows that most criteria are only weakly correlated. Two notable exceptions are"important and actionable" and "encourage better habits", and "reliable and not misleading" and "clear and pedagogical", which could be argued to be slightly redundant.

Note also that, as expected given Berkson's paradox [13], the correlations decrease if we only consider the top 10% videos on Tournesol (i.e. those that are more likely to be recommended).



(a) All videos comparisons (b) Comparisons between top 10% videos on Tournesol

Figure 6: Correlations between quality criteria

3.4 Distributions of reported comparisons

As it is not formally defined how contributors should rate a pair of videos, we expected many different expression styles. We ran a clustering algorithm (K-means) on statistics of the distribution of comparison values for each user. Figure 7 shows the typical distribution of comparison values of each of the eight clusters we identified. While some contributors provided comparisons close to "recommend equally" (cluster 3 and 4), others' comparisons were systematically towards the extreme (clusters 2, 5 and 6). This suggests that the discrepancies between their individual scores will be due to their expression style, rather than actual differences in their judgments, which justifies the research on mitigating the heterogeneity in expression styles [57, 96, 4].

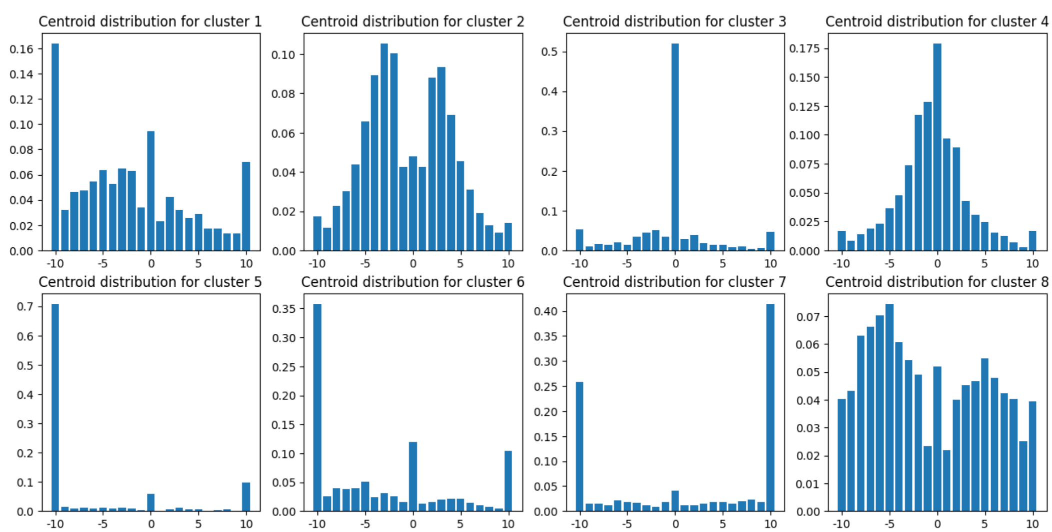
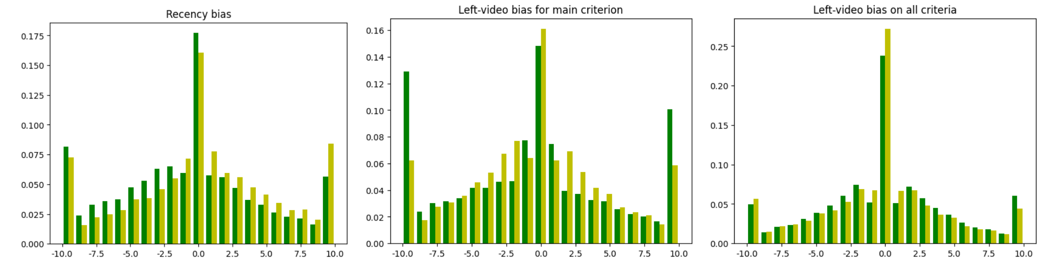


Figure 7: Example centroids of 8 clusters obtained by the K-means algorithm applied to the distributions of comparison values for each contributor with at least 20 comparisons. The clusters have sizes 144, 209, 47, 110, 23, 47, 42, 199.

3.5 Psychological biases in contributors' judgments

Our dataset exposes psychological biases in contributors' judgments. One example is a instinctive desire to over-recommend a recently watched high-quality video, known as the *recency bias* [66], which is depicted by Figure 8a. Namely, this figure plots all comparisons on the main criterion that correspond to a contributor evaluating a given video for the first time (negative scores correspond to the newly scored videos). The 95% confidence interval for the mean of first-time comparisons is $[-0.39, -0.32]$, which is arguably a surprisingly significant bias.

Another bias we observe is a tendency to favor left videos. The 95% confidence interval for the mean of the main-criterion comparisons (Figure 8b) is $[-0.54, -0.5]$. Considering all criteria (Figure 8c) yields a smaller bias, with a corresponding 95% confidence interval of $[-0.19, -0.17]$. This suggests that reflecting on more criteria reduces the left-video bias. And indeed, when they are accompanied with comparisons on other criteria, the main-criterion comparisons have a 95% confidence interval for the mean equal to $[-0.38, -0.32]$, as opposed to $[-0.67, 0.61]$ for main-criterion-only comparisons. We also observe that pretrusted contributors have a significantly reduced left-video bias (on all criteria, $[-0.05, -0.03]$ for pretrusted, $[-0.35, -0.32]$ for unpretrusted).



(a) First comparisons on main criterion (newly compared video is left). (b) Comparisons on main criterion, separated based on optional criteria. (c) Comparisons on all criteria, separated based on trust.

Figure 8: Recency and left-video biases in contributors' judgments.

3.6 Distribution of scores

Unsquashed scores (essentially, as outputs of the generalized Bradley-Terry model on contributors' comparisons) are extremely heavy tailed. Indeed, out of 791,264 scores, 4,581 deviate by more than 5 standard deviations. This is to be contrasted with the expected number 0.18 of such extreme scores, assuming a normal distribution of the scores. In fact, 428 scores deviate by more than 10 standard deviations. This observation justifies the use of comparisons to quantify the potential large deviations between top alternatives, which direct scoring approaches might fail to account for appropriately, as well as of a (robustified) quantile to standardize scores [14].

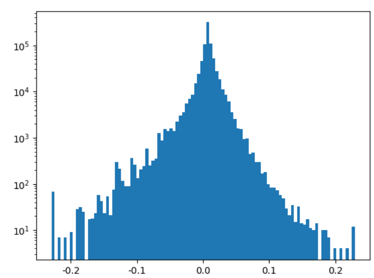


Figure 9: Distribution of unsquashed scores, with logarithmic y-scale.

4 Research challenges

Tournesol raises numerous fascinating research challenges. Below, we sketch some of these.

Aggregate the different criteria into a score. We expect the combination of many different quality criteria to yield a more reliable judgment of what content ought to be recommended at scale, or to a given specific user. However, the appropriate aggregation of our different quality criteria is still unclear, especially given probable nonlinear phenomena. How best to do this should be investigated.

Debias the contributing population. Like in many online participatory projects [10], we expect huge participation imbalances. Leveraging demographic data to debias the Tournesol recommendations, e.g., by giving more voting rights to individuals from underrepresented communities, could help, but it will require both (safely) collecting personal data and building new (secure) AIs, akin to those used by *Community Notes*[3] or by *Pol.is*[4].

Volition. As Section 3.5 highlighted, we cannot expect the Tournesol database to contain fully reliable human judgments. Many comparisons have surely been provided by contributors, at moments

[3] https://communitynotes.twitter.com/guide/en/under-the-hood/ranking-notes

[4] https://compdemocracy.org/algorithms/

when they were not paying the utmost attention to all the possible ramifications and unwanted side effects of promoting a video at scale. In particular, some judgments will arguably be more reliable than others. Such more reliable judgments are sometimes called *volitions*, rather than *preferences*. There is a need for AIs that model human psychology to distinguish between these two [54, 63].

Privacy. Tournesol's current AIs do not provide any *differential privacy* [29]. Future research should also investigate how to strengthen privacy without harming too much the quality and the security of the Tournesol scores. Perhaps most importantly, ideally, Tournesol's servers would be able to leverage private comparisons to score videos without being a single point of failure for private data protection. *Secure multi-party computations* could be a promising venue to do so [22].

Decentralize Tournesol. A longer-term goal is to fully decentralize Tournesol. In this vision, the data would no longer be stored on Tournesol's server, but would be replicated appropriately on a large number of contributors' devices. Moreover, the computations of Tournesol scores should also be decentralized, while guaranteeing *Byzantine resilience* [62]. Recent research in fully decentralized Byzantine learning has provided the building blocks of such a decentralization [32, 38], but more research is needed to understand how best to do so in the context of Tournesol.

Preference generalization. Right now, contributors are only voting on the videos that they explicitly compared. However, if they consistently voted positively all the videos of a given channel, then we could guess that they would have voted positively a new video from this channel, and to include their likely vote even when they did not compare the new video. Evidently, additional information can be leveraged to make such generalizations, such as the other video features (description, transcript, length), and the other contributors' judgments (using collaborative filtering [87]). Note however that generalization increases vulnerability risks. A careful security analysis would be required [70].

Language model alignment. Tournesol's database could help align language models, e.g. through *reinforcement learning with Tournesol feedback* [24, 79]. Determining how to combine large language models [40] with Tournesol's database to design safer models is an exciting venue for future work.

Leverage expertise. On technical topics like vaccination or climate change, especially when misconceptions are widespread in the general population, it seems desirable to assign more voting rights to experts, especially when judging the reliability of content within their domains of expertise. This issue is intimately connected to Condorcet's jury problem [25, 74].

Proof of Personhood with zero knowledge. Combatting fake accounts arguably remains the top priority to secure participatory systems. To address this, at least in democratic countries and in the short term, the state could be tasked with delivering *Proofs of Personhood* [18, 44], if possible in a zero-knowledge manner. More precisely, any citizen should ideally be able to provide to any platform a proof of citizenship, which does not enable neither the platform nor the state to identify which account is owned by which citizen. We believe that designing such a system could have applications beyond the particular case of Tournesol. Indeed, we could demand that social media only display the number of likes from users with a delivered proof of citizenship, and that their recommendation AIs be trained only by such certified users' data.

Liquid democracy Finally, future work could investigate the extent to which a liquid democracy [51] could be set up on plateforms like Tournesol. Such a system through which a contributor can delegate their votes to other voters could help combat activity bias (i.e. better accounting for inactive contributors) and expertise (if voters delegate to more competent contributors). While philosophically appealing, the security of such a system should however be first investigated [5].

5 Conclusion

This paper introduced the *Tournesol public dataset*, which is a large, secured and trustworthy database of reliable human judgments. We detailed its construction, and provided an analysis of its content. We believe that this database can help stimulate and facilitate research and development on ethical AIs, and could eventually help improve the informational diet of billions of people for the better. Given the current information crisis, we regard this as an "important and actionable" contribution.

References

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30:127–158, 2020.

[2] Gerald Albaum. The likert scale revisited. *Market Research Society. Journal.*, 39(2):1–21, 1997.

[3] Richard P Allan, Paola A Arias, Sophie Berger, Josep G Canadell, Christophe Cassou, Deliang Chen, Annalisa Cherchi, Sarah L Connors, Erika Coppola, Faye Abigail Cruz, et al. Intergovernmental panel on climate change (ipcc). summary for policymakers. In *Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, pages 3–32. Cambridge University Press, 2023.

[4] Youssef Allouah, Rachid Guerraoui, Lê-Nguyên Hoang, and Oscar Villemaud. Robust sparse voting. *CoRR*, abs/2202.08656, 2022.

[5] Shiri Alouf-Heffetz, Tanmay Inamdar, Pallavi Jain, Nimrod Talmon, and Yash More Hiren. Controlling delegations in liquid democracy. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 2624–2632. ACM, 2024.

[6] Cécile Andrzejewski. "team jorge": In the heart of a global disinformation machine. *Forbidden Stories*, 2023.

[7] Fabio Angiolillo, Martin Lundstedt, Marina Nord, and Staffan I Lindberg. State of the world 2023: democracy winning and losing at the ballot. *Democratization*, pages 1–25, 2024.

[8] Guy Aridor, Duarte Gonçalves, Ruoyan Kong, Daniel Kluver, and Joseph A. Konstan. The movielens beliefs dataset: Collecting pre-choice data for online recommender systems. In Tommaso Di Noia, Pasquale Lops, Thorsten Joachims, Katrien Verbert, Pablo Castells, Zhenhua Dong, and Ben London, editors, *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, page 1. ACM, 2024.

[9] Valentin Armhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, 567(7748):305–307, 2019.

[10] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[11] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[12] Vahid Baghi. Imdb users' ratings dataset. `https://doi.org/10.21227/br41-bd49`, December 2020. Accessed on YYYY-MM-DD.

[13] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[14] Romain Beylerian, Bérangère Colbois, Louis Faucon, Lê Nguyên Hoang, Aidan Jungo, Alain Le Noac'h, and Adrien Matissart. Tournesol: Permissionless collaborative algorithmic governance with security guarantees. *CoRR*, abs/2211.01179, 2022.

[15] Peva Blanchard, El-Mahdi El-Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 119–129, 2017.

[16] Shawn Boburg. Leaked files reveal reputation-management firm's deceptive tactics. *The Washington Post*, pages NA–NA, 2023.

[17] Maria Borge, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, and Bryan Ford. Proof-of-personhood: Redemocratizing permissionless cryptocurrencies. In *2017 IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2017, Paris, France, April 26-28, 2017*, pages 23–26. IEEE, 2017.

[18] Maria Borge, Eleftherios Kokoris-Kogias, Philipp Jovanovic, Linus Gasser, Nicolas Gailly, and Bryan Ford. Proof-of-personhood: Redemocratizing permissionless cryptocurrencies. In *2017 IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2017, Paris, France, April 26-28, 2017*, pages 23–26. IEEE, 2017.

[19] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[20] Samantha Bradshaw and Philip N Howard. The global organization of social media disinformation campaigns. *Journal of International Affairs*, 71(1.5):23–32, 2018.

[21] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. Social media and mental health. *American Economic Review*, 112(11):3660–3693, 2022.

[22] Ran Canetti, Uriel Feige, Oded Goldreich, and Moni Naor. Adaptively secure multi-party computation. In Gary L. Miller, editor, *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, USA, May 22-24, 1996*, pages 639–648. ACM, 1996.

[23] Yu Cheng, Yunzhu Pan, Jiaqi Zhang, Yongxin Ni, Aixin Sun, and Fajie Yuan. An image dataset for benchmarking recommender systems with raw pixels. In Shashi Shekhar, Vagelis Papalexakis, Jing Gao, Zhe Jiang, and Matteo Riondato, editors, *Proceedings of the 2024 SIAM International Conference on Data Mining, SDM 2024, Houston, TX, USA, April 18-20, 2024*, pages 418–426. SIAM, 2024.

[24] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017.

[25] Marie Jean Antoine Nicolas de Caritat Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale, 1785.

[26] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[27] Paresh Dave and Jeffrey Dastin. Google told its scientists to 'strike a positive tone' in ai research - documents. *Reuters*, 2023.

[28] John R Douceur. The sybil attack. In *International workshop on peer-to-peer systems*, pages 251–260. Springer, 2002.

[29] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[30] El Mahdi El Mhamdi. *Robust Distributed Learning*. PhD thesis, EPFL, 2020.

[31] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê Nguyên Hoang, and Sébastien Rouault. Collaborative learning as an agreement problem. *CoRR*, abs/2008.00742, 2020.

[32] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle. *CoRR*, abs/2008.00742, 2020.

[33] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. On the impossible safety of large AI models. *CoRR*, abs/2209.15259, 2022.

[34] El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3518–3527. PMLR, 2018.

[35] El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient learning. *CoRR*, abs/2003.00010, 2020.

[36] Julien Fageot, Sadegh Farhadkhani, Lê-Nguyên Hoang, and Oscar Villemaud. Generalized bradley-terry models for score estimation from paired comparisons. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 20379–20386. AAAI Press, 2024.

[37] Jack Farchy. Oil trader sues uae claiming smear campaign bankrupted his firm. *Bloomberg*, 2024.

[38] Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. Robust collaborative learning with linear gradient overhead. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 9761–9813. PMLR, 2023.

[39] Sadegh Farhadkhani, Rachid Guerraoui, Lê Nguyên Hoang, and Oscar Villemaud. An equivalence between data poisoning and byzantine gradient attacks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 6284–6323. PMLR, 2022.

[40] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021.

[41] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

[42] Christina Fink. Dangerous speech, anti-muslim violence, and facebook in myanmar. *Journal of International Affairs*, 71(1.5):43–52, 2018.

[43] Philippa Foot. The problem of abortion and the doctrine of double effect. *Oxford Review,*, 5, 1967.

[44] Bryan Ford. Identity and personhood in digital democracy: Evaluating inclusion, equality, security, and privacy in pseudonym parties and other proofs of personhood. *CoRR*, abs/2011.02412, 2020.

[45] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artif. Intell.*, 283:103261, 2020.

[46] Elia Gabarron, Sunday Oluwafemi Oyeyemi, and Rolf Wynn. Covid-19-related misinformation on social media: a systematic review. *Bulletin of the World Health Organization*, 99(6):455, 2021.

[47] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021.

[48] Dominique Geissler, Dominik Bär, Nicolas Pröllochs, and Stefan Feuerriegel. Russian propaganda on social media during the 2022 invasion of ukraine. *EPJ Data Science*, 12(1):35, 2023.

[49] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L. Isbell Jr., and Andrea Lockerd Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2625–2633, 2013.

[50] Gillian Kereldena Hadfield. *Rules for a flat world: why humans invented law and how to reinvent it for a complex global economy*. Oxford University Press, 2017.

[51] Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, and Manon Revel. In defense of liquid democracy. In Kevin Leyton-Brown, Jason D. Hartline, and Larry Samuelson, editors, *Proceedings of the 24th ACM Conference on Economics and Computation, EC 2023, London, United Kingdom, July 9-12, 2023*, page 852. ACM, 2023.

[52] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, 2016.

[53] Adam D Hernandez. Cambridge analytica. *Class, Race and Corporate Power*, 11(2), 2023.

[54] Lê Nguyên Hoang. Towards robust end-to-end alignment. In Huáscar Espinoza, Seán Ó hÉigeartaigh, Xiaowei Huang, José Hernández-Orallo, and Mauricio Castillo-Effen, editors, *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*, volume 2301 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[55] Lê Nguyên Hoang. Science communication desperately needs more aligned recommendation algorithms. *Frontiers in Communication*, 5:115, 2020.

[56] Le Nguyen Hoang and El Mahdi El Mhamdi. *Le fabuleux chantier: Rendre l'intelligence artificielle robustement bénéfique*. edp Sciences, 2019.

[57] Lê Nguyên Hoang, François Soumis, and Georges Zaccour. Measuring unfairness feeling in allocation problems. *Omega*, 65:138–147, 2016.

[58] Chiungjung Huang. A meta-analysis of the problematic social media use and mental health. *International Journal of Social Psychiatry*, 68(1):12–33, 2022.

[59] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *Current Journal of Applied Science and Technology*, pages 396–403, 2015.

[60] Jastra Kanjec. Facebook removed more than 15 billion fake accounts in two years, five times more than its active user base. *StockApps*, 2021.

[61] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[62] Leslie Lamport, Robert E. Shostak, and Marshall C. Pease. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, 1982.

[63] Mohamed Lechiakh and Alexandre Maurer. Volition learning: What would you prefer to prefer? In Helmut Degen and Stavroula Ntoa, editors, *Artificial Intelligence in HCI - 4th International Conference, AI-HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23-28, 2023, Proceedings, Part I*, volume 14050 of *Lecture Notes in Computer Science*, pages 555–574. Springer, 2023.

[64] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):181:1–181:35, 2019.

[65] Xiangru Lian, Binhang Yuan, Xuefeng Zhu, Yulong Wang, Yongjun He, Honghuan Wu, Lei Sun, Haodong Lyu, Chengjun Liu, Xing Dong, et al. Persia: An open, hybrid system scaling deep learning-based recommenders up to 100 trillion parameters. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3288–3298, 2022.

[66] David A Liebermann. *Learning and memory: An integrative approach*. Belmont, CA: Thomson/Wadsworth, 2004.

[67] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[68] Lucas Maystre. *Efficient Learning from Comparisons*. PhD thesis, EPFL, 2018.

[69] Lucas Maystre and Matthias Grossglauser. Just sort it! A simple and effective approach to active preference learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2344–2353. PMLR, 2017.

[70] Bhaskar Mehta and Thomas Hofmann. A survey of attack-resistant collaborative filtering algorithms. *IEEE Data Eng. Bull.*, 31(2):14–22, 2008.

[71] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The information society*, 37(1):35–45, 2021.

[72] Michael K Miller. A republic, if you can keep it: Breakdown and erosion in modern democracies. *The Journal of Politics*, 83(1):198–213, 2021.

[73] Paul Mozur. A genocide incited on facebook, with posts from myanmar's military. *The New York Times*, 15(10):2018, 2018.

[74] Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, pages 289–297, 1982.

[75] Ritesh Noothigattu, Snehalkumar (Neil) S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1587–1594. AAAI Press, 2018.

[76] Alina Oprea and Apostol Vassilev. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2023.

[77] Naomi Oreskes and Erik M Conway. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA, 2011.

[78] Aditya Pal, Abhilash Barigidad, and Abhijit Mustafi. Imdb movie reviews dataset. `https://doi.org/10.21227/zm1y-b270`, December 2020. Accessed on YYYY-MM-DD.

[79] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[80] Luca Righes, Mohammed Saeed, Gianluca Demartini, and Paolo Papotti. The community notes observatory: Can crowdsourced fact-checking be trusted in practice? In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 172–175. ACM, 2023.

[81] Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review. *Journal of Public Health*, pages 1–10, 2021.

[82] Allen L. Schirm Ronald Wasserstein and Nicole A. Lazar. Moving to a world beyond “p< 0.05”. *The American Statistician*, 73:1–19, 2019.

[83] Sébastien Rouault. *Practical Byzantine-resilient Stochastic Gradient Descent*. PhD thesis, EPFL, 2021.

[84] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[85] Alexander I. Saichev, Yannick Malevergne, and Didier Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer Science & Business Media, 2009.

[86] Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[87] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. Artif. Intell.*, 2009:421425:1–421425:19, 2009.

[88] Basu Prasad Subedi. Using likert type data in social science research: Confusion, issues and challenges. *International journal of contemporary applied sciences*, 3(2):36–49, 2016.

[89] Gan Sun, Yang Cong, Jiahua Dong, Qiang Wang, Lingjuan Lyu, and Ji Liu. Data poisoning attacks on federated machine learning. *IEEE Internet of Things Journal*, 9(13):11365–11375, 2021.

[90] Maxime Tellier. Enquête avisa partners : dans les coulisses de la sulfureuse agence d'influence soupçonnée de désinformation. *France Info*, 2023.

[91] Mariame Tighanimine. *L'affaiblissement des corps intermédiaires par les plateformes Internet. Le cas des médias et des syndicats français au moment des Gilets jaunes*. Conservatoire National des Arts et Métiers, 2019.

[92] Petter Törnberg. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42):e2207159119, 2022.

[93] Zeynep Tufekci. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press, 2017.

[94] Jean M Twenge, Jonathan Haidt, Jimmy Lozano, and Kevin M Cummins. Specification curve analysis shows that social media use is linked to poor mental health, especially among girls. *Acta psychologica*, 224:103512, 2022.

[95] Myriam Vidal Valero. Thousands of scientists are cutting back on twitter. *Nature*, 620:482–4, 2023.

[96] Jingyan Wang and Nihar B. Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 864–872. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[97] Kai Wang, Zhene Zou, Minghao Zhao, Qilin Deng, Yue Shang, Yile Liang, Runze Wu, Xudong Shen, Tangjie Lyu, and Changjie Fan. RL4RS: A real-world dataset for reinforcement learning based recommender system. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2935–2944. ACM, 2023.

[98] Gabriel Weimann and Natalie Masri. Research note: Spreading hate on tiktok. *Studies in conflict & terrorism*, 46(5):752–765, 2023.

[99] Valerie Wirtschafter and Sharanya Majumder. Future challenges for online, crowdsourced content moderation: Evidence from twitter's community notes. *Journal of Online Trust and Safety*, 2(1), Sep. 2023.

[100] Yu Xiong. Movielens 1m. `https://doi.org/10.21227/2kwp-cb23`, March 2021. Accessed on YYYY-MM-DD.

[101] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, Yu Xu, and Xiaohu Qie. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

A Datasheet for the Tournesol dataset

In this appendix, we provide a datasheet for the Tournesol dataset, based on the framework proposed by [47].

A.1 Motivation

For what purpose was the dataset created? The dataset was created to identify videos of public interest that should be recommended more largely. Additionally, we hope that the dataset will help motivate research on the ethics and security of recommendation algorithms.

Who created the dataset and on behalf of which entity? The dataset was created by the nonprofit Tournesol Association, which is based in Switzerland.

Who funded the creation of the dataset? The Tournesol Association is supporting the creation and maintenance of the dataset. It is in majority funded by crowdsourced donations, with occasional services to private companies.

A.2 Composition

What do the instances that comprise the dataset represent? The dataset contains mostly pairwise comparisons of videos by users. The dataset also contains vouches between users, authentication status, as well as processed data from this raw data.

How many instances are there in total? The dataset contains 20k users (703 pretrusted), 40k videos, 126 vouches, 204k comparisons along the main criterion and 703k comparisons along optional criteria.

Does the dataset contain all possible instances or is it a sample of instances of a larger set? The dataset contains all *public* judgments provided on the Tournesol platform.

What data does each instance consist of? Each user has a pretrust status, based on email domain Sybil resilience. Each comparison is along a criterion, and refers to a user and a pair of videos.

Is there a label or target associated with each instance? Each comparison takes a value between -10 and 10.

Is any information missing from individual instances? Yes, plenty, such as the time it took to provide an answer, whether it was provided on a phone or a desktop, or whether the contributor actually watched the compared videos.

Are relationships between individual instances made explicit? Some of them, yes, such as the contributor's identifier, or the videos that are compared.

Are there recommended data splits? Yes, comparisons are naturally split by criterion, or by users. Trusted/untrusted contributions could be split.

Are there any errors, sources of noise, or redundancies in the dataset? The comparisons come from humans, and are thus noisy, as well as potentially biased as discussed in the main part of the paper. Note that 4,446 comparisons were made before January 11, 2021, but because of a migration of the code, are dated on the January 11, 2021 week.

Is the dataset self-contained, or does it link to or otherwise rely on external sources? The dataset refers to YouTube videos, but could be analyzed without knowledge of the videos.

Does the dataset contain data that might be considered confidential? No. It was designed to be public.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening or might otherwise cause anxiety? Some poorly scored videos could be of this sort. Their content is not directly in the dataset, but the dataset points to them.

Does the dataset identify any subpopulations? Yes, trusted and untrusted contributors.

Is it possible to identify individuals, either directly or indirectly, from the dataset? Yes, especially given their public usernames.

Does the dataset contain data that might be considered sensitive in any way? Yes, indirectly, as it reveals consumption habits of contributors.

Any other comments? The individuals not only gave their consent, but the Tournesol also aims to make it clear that their provided data are used to design a democratic governance, and as such, could and should be scrutinized.

A.3 Collection process

How was the data associated with each instance acquired? Through the Tournesol platform `https://tournesol.app`.

What mechanisms or procedures were used to collect the data? Through the Tournesol comparison interface `https://tournesol.app/comparison`.

If the dataset is a sample from a larger set, what was the sampling strategy? Based on public/private settings selected by the contributor.

Who was involved in the data collection process and how were they compensated? Contributors are volunteers, most of whom are recruited through promotion in science YouTube videos. They are not compensated.

Over what timeframe was the data collected? The first data was collected in May 2020. The collection has been continuously ongoing since.

Were any ethical review processes conducted? Not by an institutional review board, as our work was done by a nonprofit association.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources? Yes, through the Tournesol platform that we designed.

Were the individuals in question notified about the data collection? Yes. They had to create a Tournesol account, to consent with the data collection, and to select whether to make their contributions public or not.

Did the individuals in question consent to the collection and use of their data? Yes.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Yes, contributors can delete their Tournesol account, which will delete their data from Tournesol's (public) dataset.

Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? Yes, we are consistently trying to make our project robustly beneficial.

A.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done? Yes. To output trust scores, as well as squashed individual and global scores.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? Yes. It is published in the Tournesol dataset.

Is the software that was used to preprocess/clean/label the data available? Yes. It is the open-source free-license Solidago python package.

A.5 Uses

Has the dataset been used for any tasks already? Yes, it is used to make content recommendations to 10k+ users.

Is there a repository that links to any or all papers or systems that use the dataset? Such papers and systems are listed in `tournesol.app/#research`.

What (other) tasks could the dataset be used for?

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Are there tasks for which the dataset should not be used? The dataset should not be used to harm individuals, communities or society.

A.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes. It is published on `api.tournesol.app/exports/all`.

How will the dataset be distributed? zip file downloadable from the website.

When will the dataset be distributed? Already is.

Will the dataset be distributed under a copyright or other intellectual property license, and/or under applicable terms of use? Yes, it is under ODC-By license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? Not to our knowledge.

A.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The Tournesol association.

How can the owner/curator/manager of the dataset be contacted? `hello@tournesol.app`

Is there an erratum? No.

Will the dataset be updated? Yes. It is weekly updated, based on Tournesol's users newly reported data.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances? No limit applies.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, the dataset is consistently updated every week, based on contributors' activity.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? The dataset is fully under the control of the Tournesol association. It is however under ODC-By license, thus any reuse is welcome, as long as attribution is appropriately provided.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution is, as explained, the publication of the datset.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explained the context in which the data is provided, and the limitations that this implies.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper dos not provide theoretical results.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The code base and the data is available online and under copyleft free license.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data is available at `https://api.tournesol.app/exports/all`, and the code is available at `https://github.com/tournesol-app/tournesol/`.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We did not provide statistical significance measures, mostly because statistical significance has been heavily criticized [82, 9]. Instead, we reported 95% confidence intervals. Note that the fact that they do not contain some "null hypothesis" is equivalent to saying that the null hypothesis has an associated p-value less than 5%. However, we believe that reporting confidence intervals is more meaningful, as it also communicates the effect size and an estimate of the uncertainty on the effect size.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: No significant compute resource is needed. The graphs were all produced on basic machines, without the need of, e.g., a GPU.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our data collection platform `https://tournesol.app` repeatedly stresses the fact that it aims to collect a public dataset of human judgments to help research. Explicit consent is asked when contributors create their account. We make it clear that the contributions should be made on a voluntarily basis, to help improve the security and ethics of recommendation algorithms.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Tournesol project is fully motivated by the desire to have a positive societal impact, by advancing the frontier of the research on the governance of recommendation algorithms. We believe that these positive impacts clearly outweigh, and by far, the potential negative societal impact, which could include, for instance, the ability of cybercrime to better organize themselves.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The dataset carefully annotates the source of the data, and contains information on the degree of authentication of the sources.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset is published by ourselves, under ODC-By license.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The dataset is documented in the paper, and a datasheet for datasets is provided in the appendix.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provided screenshots and contextualized the data collection process.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The research was conducted by a nonprofit Association, and did not involve an IRB. We discussed the main risk for participants, namely retaliation from the entities they criticize. We stress, however, that this is usually not increasing the risk, compared to what they may already be publishing on social media.