Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments

Lê-Nguyên Hoang, Louis Faucon, Aidan Jungo, Sergei Volodin, Dalia Papuc, Orfeas Liossatos, Ben Crulis, Mariame Tighanimine, Isabela Constantin, Anastasiia Kucherenko, et al.

▶ **To cite this version:**

Lê-Nguyên Hoang, Louis Faucon, Aidan Jungo, Sergei Volodin, Dalia Papuc, et al.. Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments. 2021. ⟨hal-03390514⟩

HAL Id: hal-03390514

https://hal.science/hal-03390514v1

Preprint submitted on 21 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments

Lê-Nguyên Hoang[1,2], Louis Faucon[2], Aidan Jungo[2], Sergei Volodin[2], Dalia Papuc[1,2], Orfeas Liossatos[1,2], Ben Crulis[3], Mariame Tighanimine[2,4], Isabela Constantin[2], Anastasiia Kucherenko[1,2], Alexandre Maurer[2,5], Felix Grimberg[1,2], Vlad Nitu[2,6], Chris Vossen[2], Sébastien Rouault[1,2], and El-Mahdi El-Mhamdi[2,7]

[1]IC, EPFL, Switzerland
[2]Tournesol Association, Switzerland
[3]University of Tours, France
[4]LISE, CNAM-CNRS, France
[5]UM6P, Benguerir, Morocco
[6]CNRS, INSA Lyon, France
[7]École Polytechnique, France

Abstract

Today's large-scale algorithms have become immensely influential, as they recommend and moderate the content that billions of humans are exposed to on a daily basis. These algorithms are the de-facto regulators of the information diet of billions of humans, from shaping opinions on public health information to organizing groups for social movements. This creates serious concerns, but also great opportunities to promote quality information [Hoa20, HFE21]. Addressing the concerns and seizing the opportunities is a challenging, enormous and fabulous endeavor [HE19], as intuitively appealing ideas often come with unforeseen unwanted *side effects* [EMH21], and as it requires us to think about what we truly and deeply prefer [Soa15].

To make progress, it is critical to understand how today's large-scale algorithms are built, and to determine what interventions will be most effective. Given that these algorithms rely heavily on *machine learning*, we make the following key observation: *any algorithm trained on uncontrolled data must not be trusted*. Indeed, a malicious entity could take control over the data, poison it with dangerously misleading or manipulative fabricated inputs, and thereby make the trained algorithm extremely unsafe. We thus argue that the first step towards safe and ethical large-scale algorithms must be the collection of a large, secure and trustworthy dataset of reliable human judgments.

To achieve this, we introduce *Tournesol*, an open source platform available at `https://tournesol.app`. Tournesol aims to collect a large database of human judgments on what algorithms ought to widely recommend (and what algorithms ought to stop widely recommending). In this paper, we outline the structure of the Tournesol database, the key features of the Tournesol platform and the main hurdles that must be overcome to make it a successful project. Most importantly, we argue that, if successful, Tournesol may then serve as the essential foundation for any safe and ethical large-scale algorithm.

1 Introduction

Algorithms are becoming increasingly influential. This is particularly the case of large-scale content recommendation algorithms, conversational algorithms and moderation algorithms. For example, in 2019, 70% of YouTube's one-billion hours of daily views resulted from algorithmic recommendations [Sol18]. Meanwhile, in China, Microsoft's conversational algorithm called Xiaoice is used by over 600 million humans [ZGLS20, Dur20]. Finally, Facebook had to remove 6 billion fake accounts from their platforms within just one year [FG19], mostly through algorithmic means. In this context, [HE19] argues that, unless massive financial and cognitive investments are made to ensure that such algorithms are robustly beneficial, they will inevitably have unwanted large-scale side effects through unsafe information dissemination. On the other hand, enormous opportunities in terms of quality information [Hoa20] and public health [HFE21] could be seized if algorithms were designed with careful consideration for their societal impact [VN19, SAHM20].

Current leading concerns of unethical conversational, moderation and recommendation algorithms include cyberbullying [HP10], fairness [MMS+19], discrimination [AFZ21], privacy [TCK19, CTW+20], job displacement [MT18], radicalization [ROW+20, MN20], political manipulation [WH18, Woo20], misinformation [WMCL19, Ara20], mute news (information that is drowned within the flood of information) [RRRR18], information overload [Roe19], anger [MB12], hate [SW17], geopolitical tensions [AKS21], addiction [TBB18, HS17], inability to focus [MIC+16], mental health [ESM+18], autonomous weapons [RHAV15], arms race [Gei16] and existential risk [Bos13].

The intermediate body crisis. As societies grew in size and complexity [Fla72], *intermediate bodies* have taken an increasingly central role, to mediate the interactions between individuals. Such intermediate bodies include lawmakers and states [Dur14], but also journalists, teachers and public health officials, among many others. At the heart of the role of each of these groups is information processing in order to judge what messages ought to be moderated, communicated or amplified.

As online activities grew, today's digital platforms have de-facto taken the role that was traditionally played by these intermediate bodies [Tig19, Had17]. This became particularly striking when, in 2020, the then President of the United States was banned from Twitter, Facebook, and Youtube, long before any court sentenced him for inciting violence in Capitol Hill. As another example, during the COVID-19 pandemic, digital platforms had to make decisions on health misinformation, often months before traditional intermediate bodies such as WHO, the CDC, and others. Arguably, traditional intermediate bodies currently lack the tools to play the important online role they had offline.

Tournesol's ambition is to empower intermediate bodies by building algorithms that make *scalable* judgments[1] which better represent what they would recommend to society. We hope to enable them to effectively mediate online human interactions, at the rate of billions of decisions per minute, as required by social medias.

Data is key. Importantly, today's large-scale algorithms heavily rely on *machine learning*. In other words, such algorithms leverage massive amounts of data to achieve high performance [FZS21].

[1] We recall, as in [HE19] that the very word "Algorithm" was coined after a lawyer, Alkhawarizmi, whose ambition in inventing Algebra was to help other lawyers make better judgments in courts by making decision making operations such as inheritance cases' transparent, repeatable and corrigible, through step-by-step guidelines. In the light of this remark, Algorithmic decision making should be viewed as an asset for society, rather than an obscure tool.

However, an immediate corollary of this is that such algorithms are *manipulated* by their data, especially given their current designs. If the data is full of discriminatory texts, then such "stochastic parrots" will repeat or favor such discriminatory texts [BGMS21]. What is especially concerning is that today's most influential algorithms are mostly powered by *user-generated* data downloaded from the web [SSP+13, WSM+19, WPN+19]. As a result, many of today's trained algorithms should be considered extremely unsafe. More generally, *any algorithm trained on uncontrolled data must not be trusted*.

A recent line of research on Byzantine resilient learning algorithms provides new tools to protect algorithms from poisoned data or malicious actors [BEMGS17, EMGR18, EGR20, EM20, Rou21]. However, such protections cannot perform miracles. Especially in heterogeneous environments, impossibility theorems apply [EFG+20a]. In particular, there can be no algorithmic trick to protect against a majority of maliciously crafted data sources.

More generally, the safety of learning algorithms demands the safety of their training datasets. In particular, to design trustworthy ethical algorithms, it is essential to train them on large, secured and trustworthy datasets of human ethical judgments. To the best of our knowledge, today, there is no such dataset. As a result, according to our reasoning, there can currently be no safe learning algorithm. The goal of Tournesol, available at `https://tournesol.app`, is to remedy this state of affairs, by constructing the largest, most secured and most trustworthy dataset of reliable human judgments ever created.

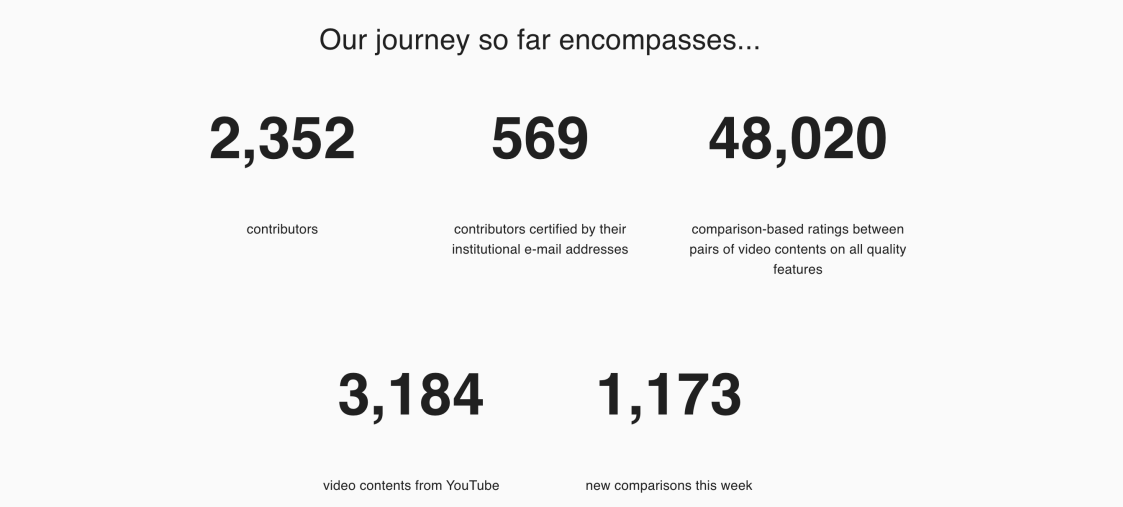


Figure 1: Global statistics of Tournesol on May 27, 2021

Introducing Tournesol. More specifically, Tournesol aims to elicit comparison-based judgments on what videos are preferable to recommend widely, and on quality features such as "reliable and not misleading", "important and actionable" and "engaging and thought-provoking", among others.

These judgments are then leveraged by a machine learning model, which infers a global score for each video. The model roughly combines the Bradley-Terry model [BT52, May18] and the Lichavi framework [FGH21], which allows to provide Byzantine resilience and (approximate) strategyproofness. Byzantine resilience means that the outputs of the model are resilient to a minority of contributors with arbitrary behaviors, while strategyproofness means that it is in strategic con-

tributors' best interests to provide judgments that match their true preferences. We stress that strategyproofness is particularly important to guarantee as much as possible the trustworthiness of our dataset.

Moreover, contributors are asked to provide personal information, which we use to assess their trustworthiness. In particular, we rely on email verification from trusted email domains, to protect Tournesol against Sybil attacks, i.e., attackers creating and having multiple identities [Dou02]. In the near future, we also plan to deploy a vouching mechanism to include more contributors without sacrificing too much security. Contributors are also asked to report their degrees and expertise, as well as demographic data.

Tournesol's interface allows contributors to either provide data publicly or privately. In the former case, their data will be appended to the Tournesol public database, which is readily downloadable from the Tournesol home page[2]. We encourage its widespread reuse and remix, given appropriate attribution, under the conditions of the license CC BY-SA[3]. Moreover, we ask that the entities who reuse the data pay great attention to strategyproofness considerations, for the trustworthiness of the Tournesol database depends on this. Finally, we ask any entity that reuses our data to do so responsibly, and, as much as possible, for the good of all of humanity. In particular, we ask them to avoid reusing our data for, e.g. advertisement targeting, especially if what is advertised is not for the good of our contributors or of humanity.

Tournesol is deeply attached to transparency. Our code is open-source[4], and nearly all of our discussions on bugs and new functionalities are publicly available[5]. Our front-end uses React [Gac15], and the back-end relies on Django [HKM09], with machine learning algorithms running with Tensorflow [ABC+16]. Our server's statistics are also public[6]. Note also that the Tournesol platform is still rapidly evolving. The present document will be updated accordingly every few months. We refer however to the Tournesol wiki, available at `https://wiki.tournesol.app`, for more frequently updated information.

Structure of the paper. The rest of the paper is organized as follows. Section 2 will present the key elements of our databases, and our motivations to collect the data we chose to collect. It will also briefly discuss our privacy policy and options. Section 3 then presents the algorithm currently in use on Tournesol. We stress in particular that our algorithm, based on Licchavi [FGH21], obeys the principle "one contributor, one unit force", which makes it somewhat strategyproof. More importantly, we hope that this section will motivate and inspire new research into analyzing user-generated comparison-based judgment datasets like Tournesol's. Section 4 presents our preliminary analysis of our dataset. Since the data collection process is ongoing, we stress that the statistics we present are bound to evolve within the next months and years. Nevertheless, we hope to thereby give insights into both the data we have collected so far, and to illustrate the kind of insights that the analysis of our database can yield. Section 5 then provides a list of research challenges that we regard as urgent to tackle for safer and more ethical algorithms in general, and for the specific case of Tournesol in particular. Finally, Section 6 provides a summary, as well as a call for contributions to make Tournesol a success, and to move towards safer and more ethical large-scale algorithms.

[2] `https://tournesol.app/tournesol.public.latest.csv.zip`
[3] `https://creativecommons.org/licenses/by-sa/2.0/`
[4] `http://github.com/tournesol-app/`
[5] `https://discord.gg/3KApy6tHAM`
[6] `https://munin.tournesol.app/`

2 The databases

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

2.1 Comparison-based judgments

To increase the quality of our data, we ask contributors to provide comparison-based judgments, by building upon a large literature on the value of such comparisons [Fes54, BT52, May18].

It is important to stress that the extent to which a piece of content should be recommended must not be measured by a binary variable. Indeed, while some pieces of content are highly undesirable to recommend at all, others are somewhat desirable to recommend widely, and others still are very important to recommend at scale. The same remark arguably holds for all the other quality criteria considered by Tournesol. A video can "encourage better habits", be "layman-friendly", and promote "diversity and inclusion" to varying degrees.

This suggests that videos should be scored on a continuous scale for each quality criterion. This is what Tournesol eventually proposes as an end product on its platform. Contributors could have been asked to judge videos on a Likert scale [Lik32]. But, while the Likert scale provides valuable information [Nor10, WTL16], the Likert scale has been widely criticized for being an unreliable measure and difficult to interpret [Alb97, JKCP15, Sub16]. In the case of Tournesol, we were concerned in particular that contributors abuse extreme values on the scale, or that different contributors use different values of the scale to mean different things [LJMZ02]. Intuitively, for instance, if a math video is consistently reviewed by extremely rigorous contributors, it might end up with a worse "reliability and not misleading" score than a video on another topic that was reviewed by much less demanding contributors. Moreover, we aim to obtain a much more fine-grained judgment of video quality.

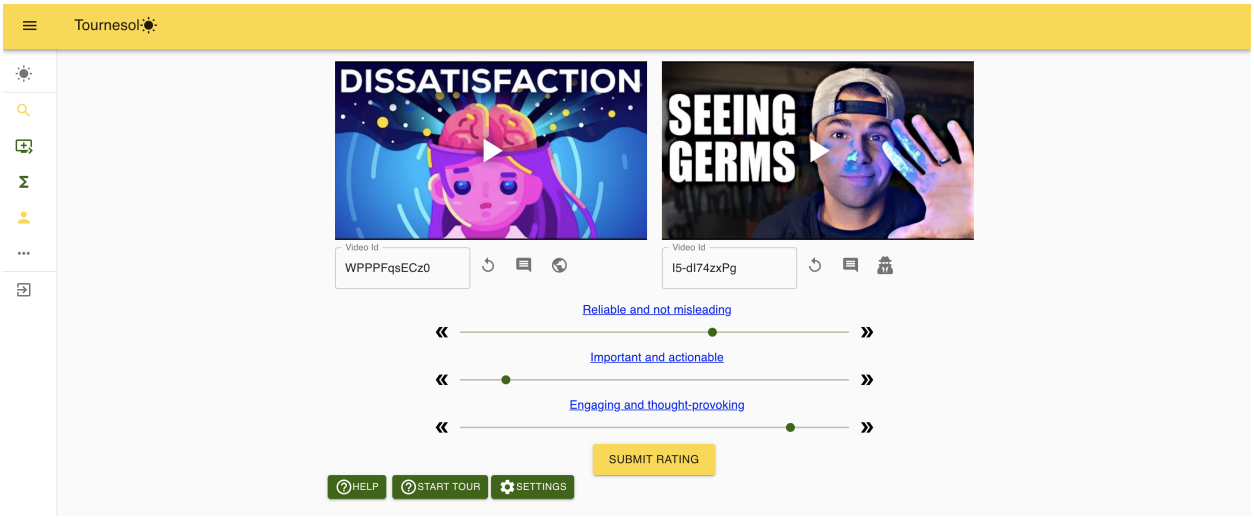


Figure 2: This is the Tournesol interface through which contributors are asked to provide judgments. The judgments are comparisons of different video content along different quality criteria.

To circumvent the difficulty of extreme scoring and of incomparable individual scales, Tournesol asks contributors to provide comparison-based ratings (see Figure 2). 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 continuous scale, from 0 to 100. When rating a quality criterion Q, the value 0 means that the contributor judges that the left video content is far better in terms of Q, while the value 100 means that they believe the right video content is far better in terms of Q. Evidently, if the contributor reports 50, it means that they put the slider in the middle, as they judge that the two videos should be scored similarly according to Q.

We believe that such comparative judgments will yield more reliable human judgments. We hope that this will allow for the design of more reliable algorithm training.

2.2 Quality criteria

Tournesol currently proposes ten *quality criteria* to rate, only one of which is available by default.

The only *default* quality criterion is "Should be largely recommended". Tournesol chose to single out this one criterion as it is arguably the bottom line criterion for constructing recommendation algorithms. Moreover, Tournesol chose to make it the only default quality criterion to facilitate the use of Tournesol for a large public of contributors. We were particularly concerned about the risk of overwhelming new contributors with too many complex queries.

Nevertheless, Tournesol provides contributors with the possibility to rate nine other *optional* quality criteria:

Reliable and not misleading: Content that scores high on 'reliable and not misleading' should make nearly all viewers improve their global world model, despite viewers' biases and motivated reasoning.

Important and actionable: Content that scores high on 'important and actionable' should present data and arguments with major consequences, as well as actionable plans that would have a large impact.

Engaging and thought-provoking: Content that scores high on 'engaging and thought-provoking' should catch the attention of a larger audience, trigger their curiosity and a desire to find out more or question their own beliefs.

Encourages better habits: Content that scores high on 'encourages better habits' should be successful at motivating viewers to grow and improve themselves.

Clear and pedagogical: Content that scores high on 'clear and pedagogical' should help viewers understand all the elements that lead to a conclusion.

Layman-friendly: Content that scores high on 'layman-friendly' should be accessible to a very large audience.

Diversity and inclusion: Content that scores high on 'diversity and inclusion' should celebrate diversity and be appealing to minority groups.

Resilience to backfiring risks: Content that scores high on 'resilience to backfiring risks' should be safe to recommend to all sorts of viewers, with limited risks of misunderstandings.

Entertaining and relaxing: Content that scores high on 'entertaining and relaxing' should entertain and relax viewers.

While detailed descriptions of all the criteria are provided on the Tournesol wiki, at `https://wiki.tournesol.app/index.php/Quality_criteria`, data scientists and researchers should probably not expect most contributors to read thoroughly our descriptions. Arguably, most contributors will more likely judge these criteria according to their own understanding, which will be mostly based on the name of the criteria.

Note also that contributors can skip the judgment of a quality criterion. The database is partially sparse in this regard.

2.3 Confidence

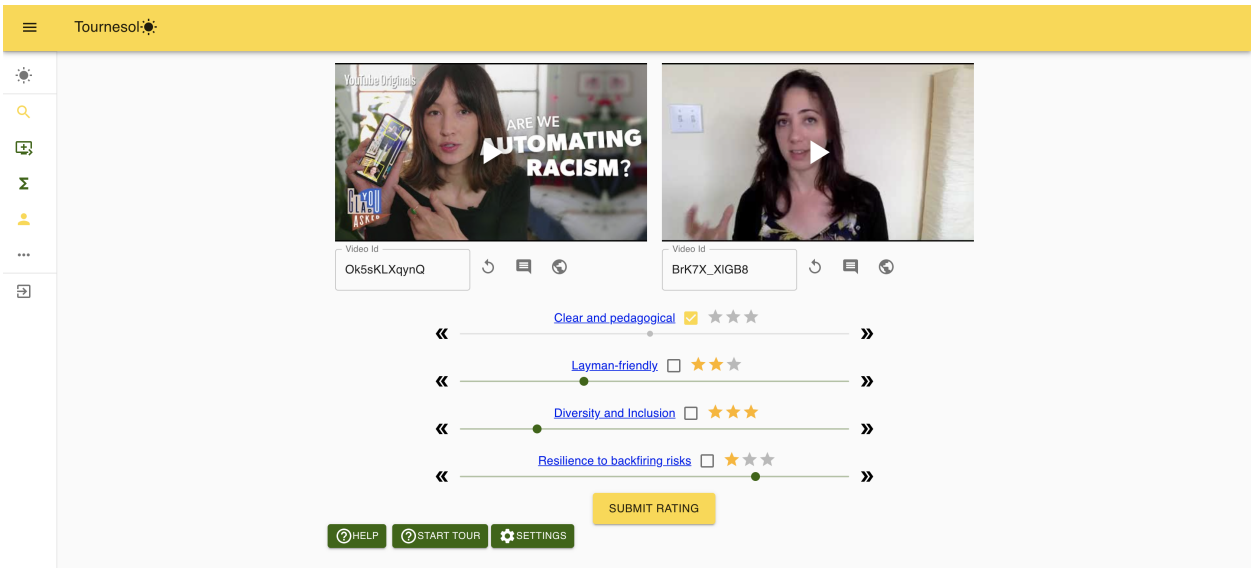


Figure 3: Tournesol allows contributors to judge videos, as well as the confidence they have in their judgments

Tournesol also proposes the option for contributors to assess their confidence in their ratings along each quality criterion, on a scale from 0 to 3, as illustrated in Figure 3. A confidence of 0 is equivalent to skipping the criterion altogether.

Such data are currently used by our learning algorithm to determine both the contributors' scores, and their impacts on the global Tournesol scores.

2.4 Email verification

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* [BKJ+17] to verify each active Tournesol account, and to thereby prevent *Sybil attacks* [Dou02]. 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 email domain.

Clearly, this excludes email provider domains like @gmail.com and personal domains like @my-personal-website.com. Indeed, the concern is not only that the email domain owner 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/email_domains`. It includes domains like @epfl.ch, @who.int and @rsf.org.

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.

2.5 Vouching mechanism

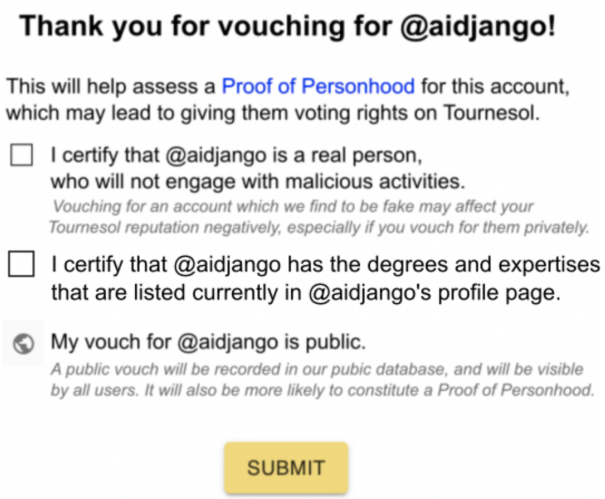


Figure 4: Any account on Tournesol can vouch for another account. This helps Tournesol verify more accounts in a secured manner.

Tournesol also proposes a vouching mechanism (see Figure 4). 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. Each email-verified account is then given a certain amount of vouching power, and each account must receive a certain amount of vouching (weighted by vouching power) to be certified. We refer to `https://wiki.tournesol.app/index.php/Vouching_mechanism` for more (updated) details.

We are currently investigating the design of a reliable, Byzantine-resilient and scalable vouching mechanism (see Section 5.2).

2.6 Meta-data

To better understand our contributors' rating patterns, and identify which ratings follow from thoughtful reflections, Tournesol measures the contributors' response times. Tournesol also records

the motions of the sliders. We believe this to be particular important for volition learning (see Section 5.8).

2.7 Privacy

Contributors can provide ratings 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; namely, only Tournesol's server will have access to such data. Conversely, all comparisons that involve two publicly rated videos are public, and can be downloaded from the Tournesol home page.

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 algorithms. However, we acknowledge that, because of social and political pressures, some judgments are dangerous to make public. In such cases, Tournesol allows contributors to provide these judgments in a private way that nevertheless affects the Tournesol global scores, and thus what will be recommended by our platform.

2.8 Personal information

Each Tournesol contributor has a dedicated Tournesol contributor page, where they can access their individual scores for different videos, as well as statistics based on their recommendations.

On this page, we also ask contributors to provide personal information, publicly or privately. Again, we encourage transparency from our contributors, as this will greatly facilitate research and the design of solutions for safer and more ethical algorithms. But we acknowledge that this comes at a cost for our contributors, which they may want to avoid.

Personal information of great interest to Tournesol includes the contributor's expertise and degrees, as well as their demographic data. Such data are particularly critical for Tournesol to understand its sampling bias, and to design debiasing solutions to avoid discriminatory algorithms.

3 Tournesol's learning algorithm

In this section, we briefly describe the learning algorithm used by Tournesol to transform comparison-based ratings into individual and global scores. We hope that this will increase the transparency of Tournesol, inspire further research into leveraging our data and stress the importance we assign to strategyproof learning. Figure 5 lists the current top recommendations on Tournesol.

Note that each quality criterion is treated independently. Thus, without loss of generality, we only focus on a single criterion here.

3.1 Licchavi

Tournesol uses the Licchavi framework [FGH21]. More precisely, denote $[N] = \{1, \dots, N\}$ the set of verified Tournesol contributors, and $[V] = \{1, \dots, V\}$ the set of rated videos. We denote $\mathcal{D}_n = \{(v, w, r)\}$ a set of ratings by contributor $n \in N$. We assume that $r = (slider - 50)/50 \in [-1, 1]$, where $slider \in [0, 100]$ is the position of the slider discussed in Section 2.1. The variables θ_{nv} correspond to the (learned) score of video v given by contributor n, while ρ_v will be the global score of video v. Finally, for clarity, we denote $\vec{\mathcal{D}}$ and $\vec{\theta}$ the tuple of data $\mathcal{D}_n$ and θ_n.

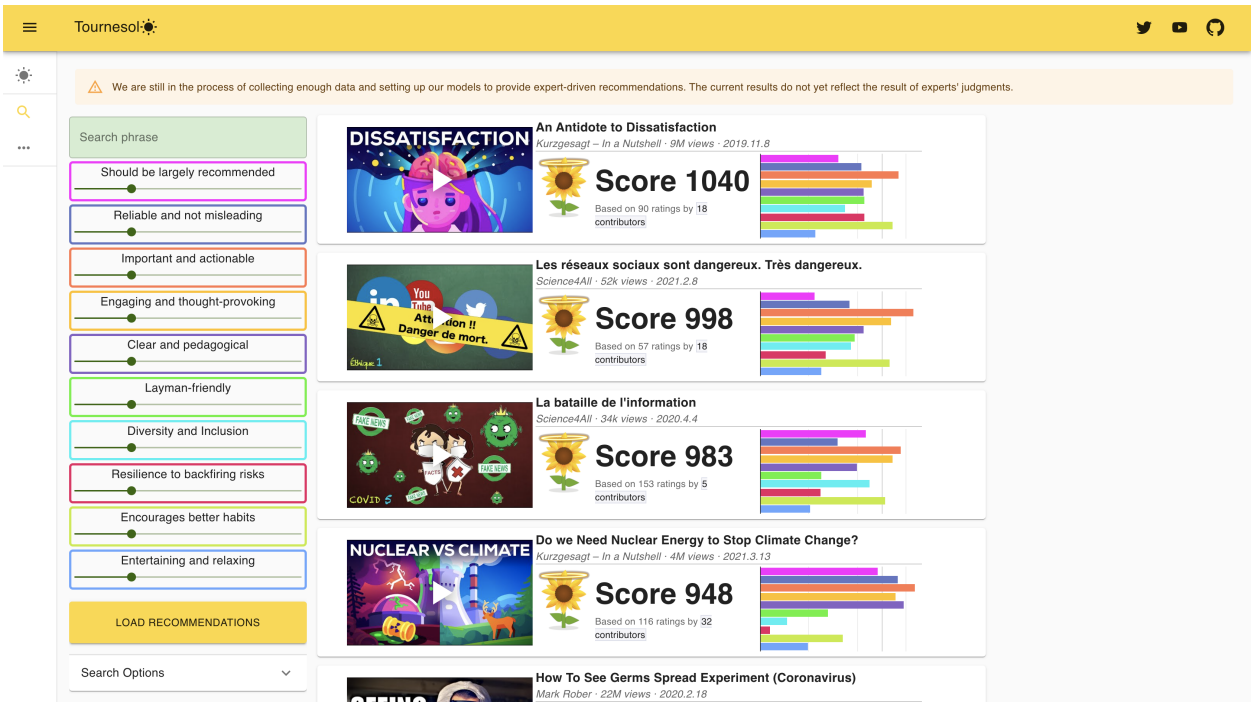


Figure 5: Top recommendations on May 27, 2021, on Tournesol's website. Note that, on the Tournesol website, users can customize their recommendations by adjusting the importance they assign to the different quality criteria.

Tournesol then considers the following loss function:

$$\text{Loss}(\rho, \vec{\theta}, \vec{\mathcal{D}}) \triangleq \sum_{n\in[N]} \sum_{(v,w,r)\in\mathcal{D}_n} \ell(\theta_{nv} - \theta_{nw}, r) + \lambda \sum_{n\in[N]} \sum_{v\in[V]} w_{nv} \left|\theta_{nv} - \rho_v\right| + \nu\lambda \sum_{v\in[V]} \rho_v^2, \qquad (1)$$

where ℓ is a loss-per-input function which will be detailed in the next section, λ is the weight of collaboration, w_{nv} is the weight of contributor n on the score of video v, and ν is the weight of the regularization of the global scores. A precise analysis of this loss function is beyond the scope of this paper, and is currently being investigated.

Tournesol's individual scores θ_{nv} and global scores ρ_v are then computed by minimizing the Tournesol loss function. In particular, the global scores are used to make recommendations (see Figure 6) and are displayed on Youtube, for users using on browser extension (see Figure 7).

3.2 Bradley-Terry model

Let us now detail the loss-per-input function ℓ currently used by Tournesol. We derive this loss from the classical Bradley-Terry model [BT52], which assumes that video comparisons are binary: either the video v is better than w (expressed by a rating $r = 1$), or vice versa ($r = -1$). The probability of $r = 1$ (i.e., preferring video v to video w) depends on the difference t between the videos' scores: $t \triangleq \theta_{nv} - \theta_{nw}$. The larger the difference t is, the more probable it is that the contributor declares preferring v to w. The loss is then the negative likelihood of the data, assuming independence of different ratings, i.e.

$$\ell(t, r) \triangleq \ln\left(1 + \exp(tr)\right). \qquad (2)$$

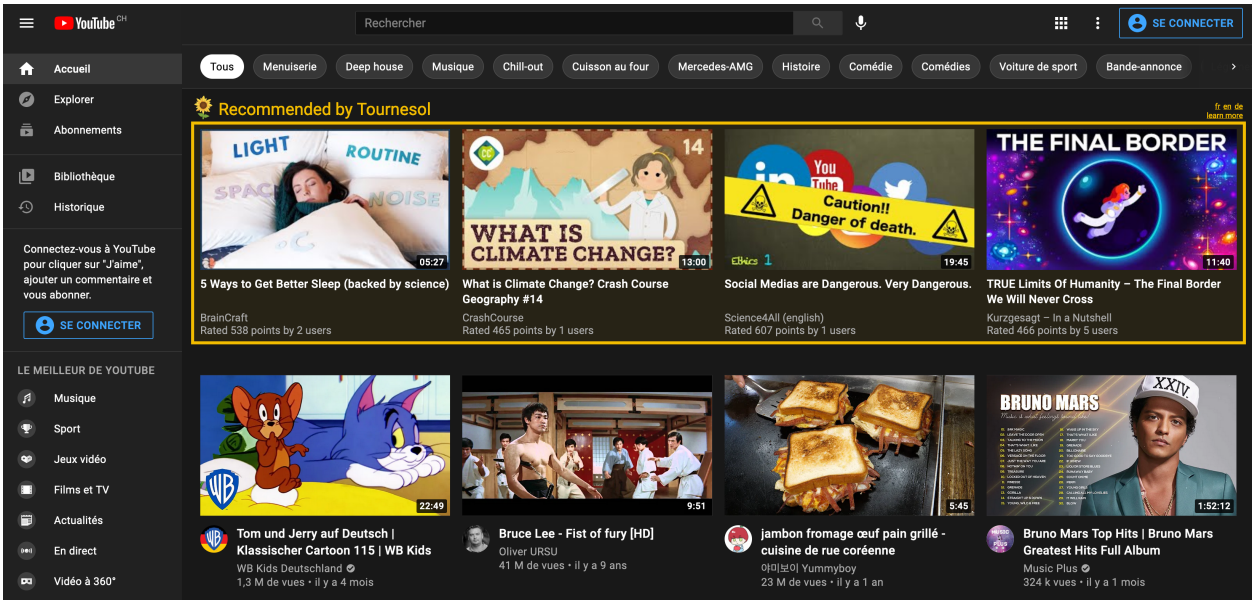


Figure 6: Using our Chrome or Firefox extensions, Tournesol users are provided with Tournesol recommendations, based on the global scores ρ^*.

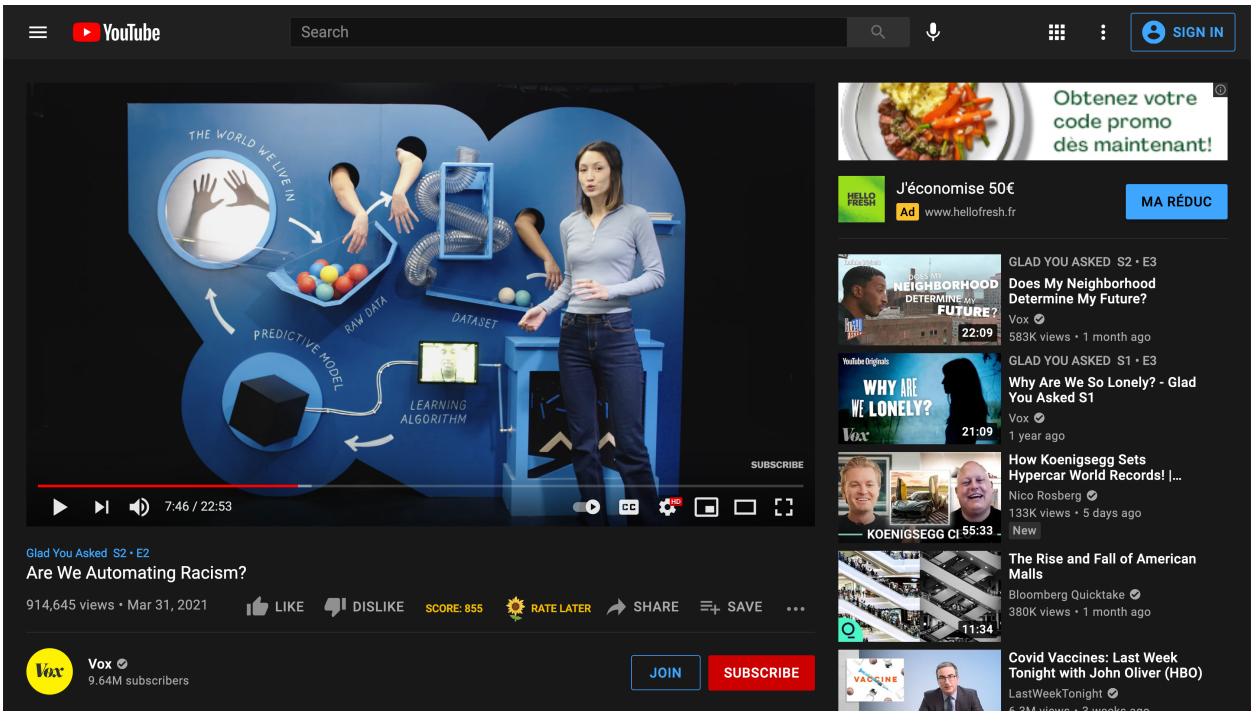


Figure 7: Using our Chrome or Firefox extensions, Tournesol users are provided with Tournesol's global scores ρ^* on the video page, next to the like/dislike statistics. Contributors can also click on "rate later" to add the video to their Tournesol rate-later list.

Note that this loss is also essentially the logistic regression loss.

The usual interpretation of this loss is that v will be one point above w (i.e., $t = 1$), if it is e times more likely to be rated better. However, this model is no longer adequate when contributors provide continuous ratings $r \in (-1, 1)$. In fact, with the Bradley-Terry model, providing a rating $r = 0$ is equivalent to not contributing at all. However, a rating $r = 0$ may be more accurately interpreted as the contributor saying that the two videos should have similar scores. The Bradley-Terry model fails to capture the nuanced information contained within continuous ratings.

This remark calls for future research to investigate alternative learning algorithms, which may yield better interpretability, and may better capture the intuitive idea that, when a contributor judges $r \approx 0$, they are actually saying that the two compared videos should have similar scores.

3.3 Hyperparameters

We ran preliminary experiments to understand the impact of the hyperparameters of our learning model. Intuitively, λ measures how much the global scores are used to adjust each contributor's local scores. Large values of λ mean that the global scores are regarded as a very reliable prior. Smaller values of λ imply that the contributor's scores are likely to quickly diverge from the global scores. In fact, $1/\lambda$ is the order of magnitude of the prior standard deviation on contributor n's score for video v, given the common score ρ_v. Intuitively, the more diversity there is in contributor's judgments, the smaller the value of λ should be. Future research will address how to dynamically adjust λ.

The hyperparameter ν weighs the prior regularization on the scores. More importantly, its value determines the Byzantine resilience of our learning model [FGH21]. Larger values of ν imply that a single contributor can hardly affect the global scores. As the number of contributors on Tournesol grows, we plan to decrease the value of ν.

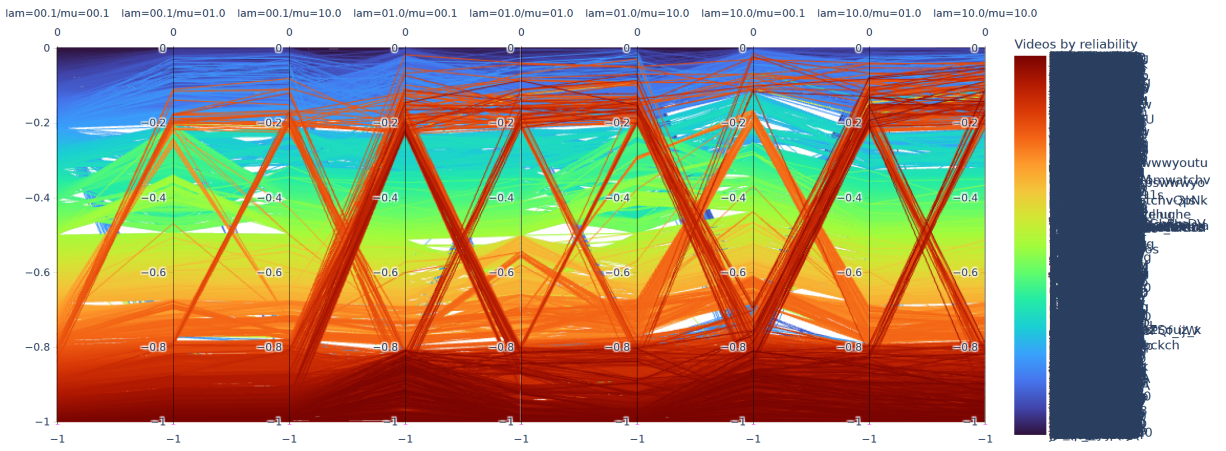


Figure 8: Video ranking of different videos based on the values of the hyperparameters λ and $\mu \triangleq \nu\lambda$.

Figure 8 shows how sensitive the scores of the videos are with respect to the hyperparameters. This should not be surprising, given the heterogeneity of how often different videos have been rated.

In particular, a when $\nu = \mu/\nu$ is small, then the rating of a video cannot be large if the video was not rated by many contributors. Further investigations of the impact of hyperparameters, and what a safe and ethical choice of hyperparameters should be, are needed.

Note however, that the effect of the hyperparameter ν vanishes in the limit of a large number of contributors. At this point, the global scores are simply the medians of the contributors' scores. Likewise, the effect of the hyperparameter λ also vanishes in the limit of a large number of data per contributor. In other words, if Tournesol is used by a large number of very active contributors, then the choice of the hyperparameters hardly matters.

Currently, the platform uses the values $\lambda = \nu = 1$. Finally, we defined $w_{nv} \triangleq R_{nv}/(C + R_{nv})$, where R_{nv} is the number of times contributor n rated video v against other videos. This allows to favor contributors who provided more ratings to a video. We set the hyperparameter $C = 3$. As a result, a contributor would obtain half of their maximal influence by rating a video three times, and 75% of it by rating it nine times.

3.4 Non-verified contributors

Note that, given that ratings are sparse, inferring individual scores from individual ratings only is suboptimal. In particular, this approach is vulnerable to Stein's paradox [Ste56, JS61]. Instead, to learn individual scores, even for non-verified contributors, we leverage the learned global scores ρ^* (note that minimizing Loss does this automatically for verified contributors). This corresponds to minimizing this loss for every non-verified contributor n:

$$\text{Loss}_n(\theta_n, \mathcal{D}_n) \triangleq \sum_{(v,w,r)\in\mathcal{D}_n} \ell(\theta_{nv} - \theta_{nw}, r) + \lambda \sum_{v\in[V]} w_{nv} \left|\theta_{nv} - \rho_v\right|. \tag{3}$$

Such learned scores are used to provide individual recommendations on the non-verified contributors' Tournesol page.

3.5 Customizable recommendations

Tournesol applies the learning algorithm described above to the different quality criteria. This allows users to obtain customizable recommendations and search results, by adjusting the importance they assign to the different quality criteria. The videos are then ranked based on the weighted sum of the scores per criteria, where the weights are given by the positions of the user's sliders (see Figure 5). However, we leave open the question of designing personalized robustly beneficial content recommendation, as discussed in Section 5.14.

3.6 Strategyproofness

We stress that the Licchavi framework with such ℓ_1 collaborative regularizations has been tied with strategyproofness theorems in [FGH21]. We acknowledge, however, that the strategyproofness theorem in [FGH21] does not quite apply to our setting though, partly because our loss-per-inputs are not actually gradient-PAC, but more importantly because our queries are not coordinate-wise. Nevertheless, our framework applies the fundamental fairness principle "one voter, one unit force" proposed by [EMFGH21], which suggest a reasonable amount of strategyproofness (and of Byzantine resilience). Future research is however needed to better determine the extent to which our algorithms are robust and strategyproof.

4 Statistics

To highlight the value of our database, we present insightful preliminary data analyses. Evidently, our data will evolve, hopefully drastically, in the coming months.

4.1 Contributors' contributions

Figure 9 displays the number of contributions per user among the top current contributors. Perhaps unsurprisingly, this statistics seems to follow a heavy tail distribution, with a few contributors providing most of the ratings, and the majority of them providing very few ratings. Fortunately, our learning algorithm, based on Licchavi with norm-based collaborative regularizations, fits the principle "one voter, one unit force" [EMFGH21], and thus prevents the leading contributors from having an uncontrollable influence on global scores.

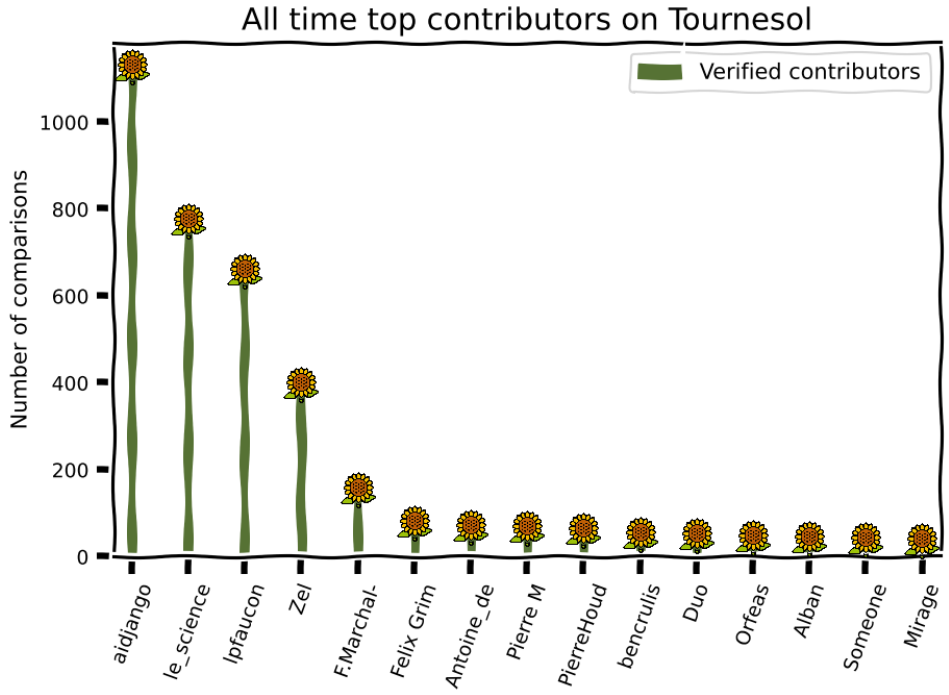


Figure 9: This figure displays the top contributors on Tournesol, and their number of ratings.

4.2 Connectivity of the data

The left part of Figure 10 shows the connectivity of the graph of video comparisons. In this graph, two videos are connected if they have been compared at least once on the Tournesol platform. While the graph features a dense center, many videos are only weakly connected to most other videos. The connectivity of the graph is important, as it allows to guarantee that all videos are scored along the same scale.

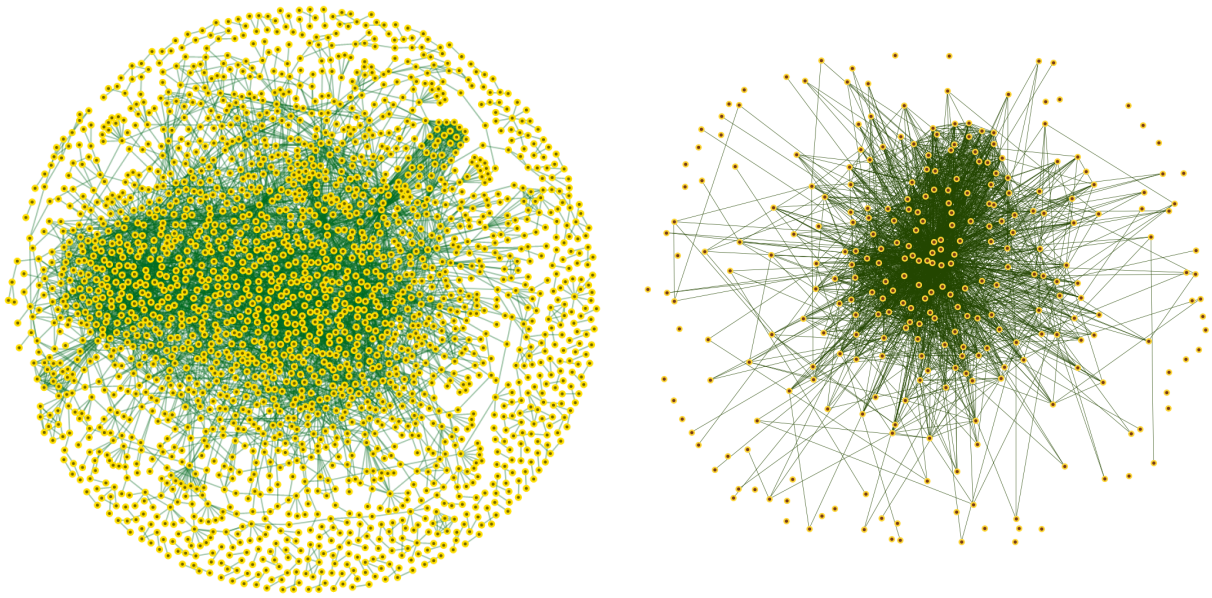


Figure 10: On the left, each sunflower on the figure represents a video, and two videos are connected if there is a contributor who compared them at least once. On the right, each sunflower on the figure represents a contributor, and two contributors are connected if there is a video that they both rated.

The right part of Figure 10 displays the commonalities between contributors. Namely, nodes on the graph are contributors, and two contributors are connected if their lists of rated videos have a non-empty intersection.

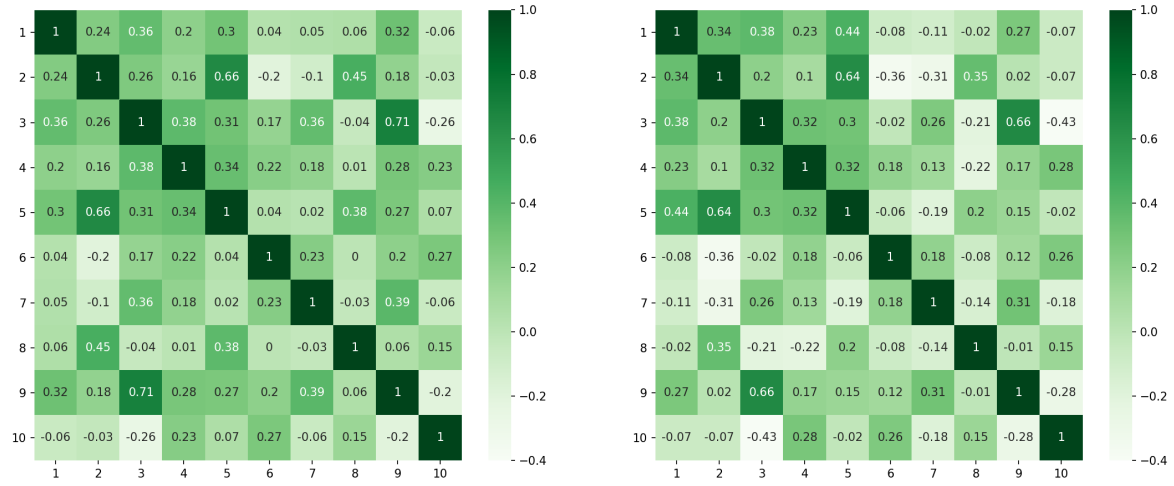
4.3 Correlations between criteria

Figure 11a reports the correlations between quality criteria. It shows that the choice of our criteria is somewhat reasonable, as most criteria are only weakly correlated. Some criteria are however somewhat redundant like "reliable and not misleading", "clear and pedagogical" and "resilient to backfiring risks". Similarly, "important and actionable" and "encourage better habits" are also quite correlated.

As expected given Berkson's paradox [Ber46], the correlations decrease if we only consider the top 10% videos on Tournesol (Figure 11b). This is important as these are the videos that Tournesol users will be more exposed to.

4.4 Rating patterns

As it is not formally defined how contributors should rate a pair of videos, we expected many different rating styles. Figure 12 shows how two contributors whose rating patterns are drastically different, despite rating similar videos. Namely, while contributor "aidjango" provided ratings close to "indifferent", contributor "le_science4all" provided more extreme ratings. This suggests that the discrepancies between their individual scores will be due to their rating style, rather than actual differences in their judgments. As discussed in Section 5.4, research should address how to adapt



(a) All videos rated on Tournesol

(b) Top 10% videos rated on Tournesol

Figure 11: Correlations between quality criteria, whose numbering are given in Table 1. Ideally, a set of criteria should be orthogonal enough to provide complementary information with minimal effort on the contributor's side.

#	Name of criteria
1	Should be largely recommended
2	Reliable and not misleading
3	Important and actionable
4	Engaging and thought-provoking
5	Clear and pedagogical
6	Layman-friendly
7	Diversity and Inclusion
8	Resilience to backfiring risks
9	Encourages better habits
10	Entertaining and relaxing

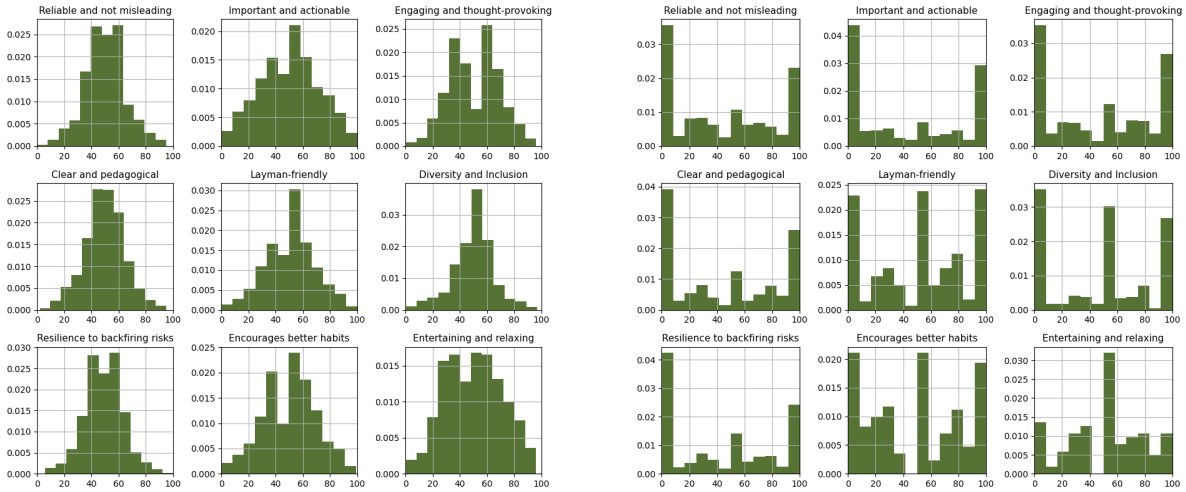
Table 1: Numbering of the criteria used in Figure 11.

the learning model to contributors' different rating patterns.

4.5 Pareto front

Figure 13 shows the number of videos per Pareto rank. To recall, the Pareto rank of a video is the number of videos that need to be removed so that the video becomes a Pareto-optimal video, i.e., a video that no other video outperforms on all quality criteria.

We note that, currently, there is a large number of pareto-optimal videos (> 50) because there are 10 dimensions allowing for plenty of trade-offs. This highlights the usefulness of customizable recommendations. It is also noteworthy that many videos have rank 3 or less. This is likely due to the fact that current contributors mostly rate videos that they deem particularly worthy of recommendation. It will be interesting to see how these statistics evolve as our database grows,



(a) Contributor "aidjango" (b) Contributor "le_science4all"

Figure 12: Distribution of the ratings for each criteria, showing difference of rating patterns between two Tournesol contributors.

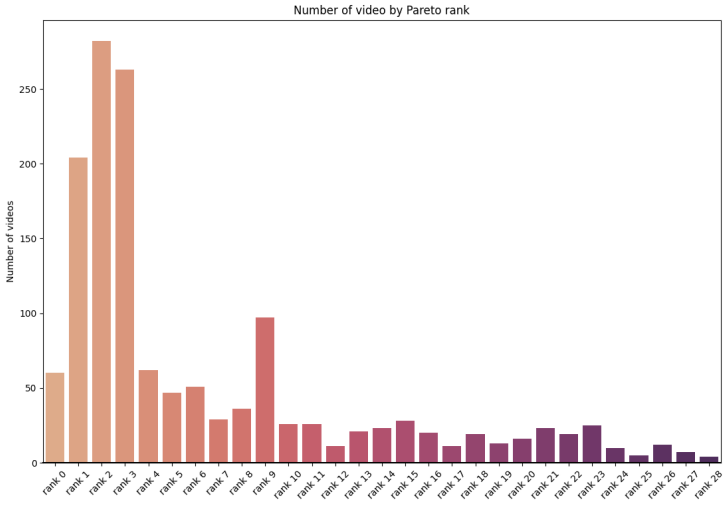


Figure 13: Number of videos per Pareto rank.

and contains a more diverse range of video rating styles.

5 Research challenges

Tournesol raises numerous fascinating research challenges. Below, we sketch some of these challenges. We would welcome any potential collaborations on any of the following challenges, and actively encourage the use of our public database for good, and, in particular, for research pur-

poses.

5.1 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 a specific user. However, the appropriate aggregation of our different quality criteria is still unclear, especially given probable nonlinear phenomena. For instance, a reasonably reliable content seems vastly superior to a misleading content, whereas an extremely reliable content does not seem vastly superior to a reasonably reliable content. We hope to investigate how best to do this, by comparing aggregations of optional quality criteria to the "Should be largely recommended" criterion, and by factoring in discrepancies between different contributors' ratings, given their expertise and contributing profiles.

5.2 Verifying more contributors

We hope to leverage vouching to verify more contributors in a Byzantine-resilient manner. However, it is unclear so far what are the best algorithms to achieve this. We are currently investigating this relatively general problem (see Section 2.5), which is tightly connected to ideas such as *Web of Trust* [Mue21] or reputation mechanisms [dORM+20].

5.3 Debiaising the contributing population

Unfortunately, like in many online participatory projects [ADK+18], we expect huge participation imbalances. Young male technology-savvy individuals are probably more likely to participate in Tournesol, which means that they will be overrepresented in the Tournesol database. We are currently investigating how to leverage our demographic data to debias the Tournesol recommendations, e.g., by giving stronger voting rights to individuals whose communities are underrepresented in the Tournesol database.

5.4 Learning and correcting rating patterns

The current "binomial Bradley-Terry" model assumes that all contributors rate in the same manner, given implicit score differences. In practice, as discussed in Section 4.4 and as depicted by Figure 12, different contributors will have very different rating habits. To make sure that the learned individual scores are fairly comparable, the rating model should be hyperparameterized, and the hyperparameters should be learned and adjusted per contributor. We are currently investigating ways to do this.

5.5 Leveraging 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 [Con85, NP82]. We are currently investigating how best to leverage personal information to determine appropriate voting rights.

5.6 Understanding and improving contributors' experiences

To understand the reliability of Tournesol's data, it is critical to understand the psychology of contributors when they provide judgments. We are currently investigating contributors' thought processes, and how rating on Tournesol affects contributors' reasonings. This work is inspired by [LKK+19], who interviewed participants of their participatory ethical algorithm design process, and found out that, by the participants' own assessments, this enabled them to improve their ethical judgments and to increase their trust in the participatory system.

5.7 Active learning

To increase the quality and the quantity of the database, it is crucial for Tournesol to help contributors select the videos to compare in the best possible ways. This is a challenging task, as it encompasses several considerations. On one hand, the comparison should provide valuable data, by typically querying interesting comparisons, involving videos with too few ratings or better connecting the graph of pairwise video comparisons. On the other hand, it is important that this demands as little cognitive investment from the contributor as possible. This typically requires suggesting the contributor to compare videos that they have recently watched and assessed. Finding out how to best do this is a research challenge that we hope to investigate soon.

5.8 Volition learning

Despite our efforts, we cannot expect the Tournesol database to contain fully reliable human judgments. We expect many of ratings will be provided by contributors who might not be considering all the possible ramifications and unwanted side effects of promoting a video content at scale when they provided their judgments. In particular, some judgments will arguably be more reliable than others. Understanding which judgments are more reliable than others is a vast research program that is critical to design safer and more ethical algorithms. More reliable judgments are sometimes called *volitions*, rather than *preferences*. Volitions are also often described as second-order preferences: they correspond to what we would prefer to prefer, rather than what we instinctively prefer [Tar10, HE19]. We hope to initiate the research on *volition learning*, by leveraging the meta-data of our database (see Section 2.6).

5.9 Privacy-preserving learning

While we believe that they have a very reasonable amount of privacy protection, our current algorithms require a deeper analysis to understand the extent to which they guarantee the protection of private ratings. Future research should also investigate how to strengthen privacy without harming too much the quality and the security of the Tournesol scores, and how to leverage private personal information about our contributors (which are currently not used) in a privacy-preserving manner. Perhaps most importantly, ideally, Tournesol would be able to leverage private ratings to score videos without being a single point of failure for private data protection. Moving forwards, a challenging research avenue is to investigate the extent to which the mission of Tournesol can be achieved without any transfer of private information.

5.10 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. Recent research in fully decentralized Byzantine learning has provided the building blocks of such a decentralization [EFG+20b], but more research is needed to understand how to best do so in the context of Tournesol.

5.11 Generalizing learning

Right now, Tournesol only leverages the Tournesol database to compute individual and global scores. However, in the longer-term, it seems desirable to leverage additional information, such as the videos' channels or their descriptions, to generalize the individual scores of a contributor to videos that they have not rated (and not even watched!). In particular, the use of natural language processing on video captions can be a promising avenue, though this probably requires a larger database than the current database.

5.12 Enable multiple individual learning algorithms

Right now, Tournesol is imposing each contributor to use our model, and in particular the binomial Bradley-Terry loss (see Section 3.2). However, especially as we move to algorithms able to generalize to videos the contributor has not watched yet, it seems desirable to enable the contributor to choose which learning model they want to use. To achieve this, Licchavi must be adapted to enable the collaboration of different local learning models.

5.13 Ethical language models

In the long run, we hope that Tournesol's database will be useful to design more sophisticated ethical algorithmic products, such as a ethical language models [BGMS21]. Typically, when prompted, such language models should consistently produce texts that are robustly beneficial to communicate, either at scale, or to targeted consumers. Determining how to combine large language models [FZS21] with Tournesol's database to design safe and ethical language models is currently a very open research challenge.

5.14 Leveraging Tournesol's scores for alignment

For most large-scale algorithmic systems, such as recommendation algorithms, the objective function is the main avenue we have to make them robustly beneficial. The problem of making this objective function safe and desirable to optimize is called the *alignment problem* [Yud16, Rus19, Hoa19]. While a large, secured and trustworthy database of reliable human judgments seems critical to solve alignment, it is unclear so far how to best leverage the Tournesol database to achieve this, especially while being personalized [FGH21] and while being robust to Goodhart's law [EMH21]. This is arguably the most challenging and the most exciting of all the problems Tournesol raises.

6 Conclusion

Summary. In this paper, we introduced Tournesol, a platform whose goal is to collect and curate a large, secured and trustworthy database of reliable human judgments. We discussed the main features of the platform, the algorithm we use today to leverage this database to make more robustly beneficial recommendations, and the numerous research challenges that must be overcome to make Tournesol a success. We strongly believe that the creation of such a database will stimulate research and development on ethical algorithms, which should help improve the informational diet of billions of people for the better. To achieve this ambitious goal, however, we will need support.

Call for action. To date, many individuals have voluntarily contributed to our code base and our database, for which we are immensely grateful. However, we still need a large amount of software development to make our platform more contributor-friendly, more secure and more scalable. For this reason, the Tournesol Association is currently searching for funding to hire top developers to maintain and develop the platform. Just as critically, Tournesol needs to attract a large number of contributors and to obtain human judgments from a diverse population. We would also be hugely grateful for the support of fellow researchers and institutions. To help us, try out the platform and please spread the word! Your contributions will be improving the quality of online discourse for everyone.

References

[ABC+16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[ADK+18] 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.

[AFZ21] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *CoRR*, abs/2101.05783, 2021.

[AKS21] Julia Alexander, Jacob Kastrenakes, and Bijan Stephen. How facebook, twitch, and youtube are handling live streams of the capitol mob attack. *The Verge*, 2021.

[Alb97] Gerald Albaum. The likert scale revisited. *Market Research Society. Journal.*, 39(2):1–21, 1997.

[Ara20] Sinan Aral. *The hype machine: How social media disrupts our elections, our economy and our health-and how we must adapt.* Currency, 2020.

[BEMGS17] 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.

[Ber46] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[BGMS21] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM, 2021.

[BKJ+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.

[Bos13] Nick Bostrom. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31, 2013.

[BT52] 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.

[Con85] 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.

[CTW+20] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *CoRR*, abs/2012.07805, 2020.

[dORM+20] Marcela T. de Oliveira, Lúcio Henrik A. Reis, Dianne S. V. Medeiros, Ricardo Campanha Carrano, Sílvia D. Olabarriaga, and Diogo M. F. Mattos. Blockchain reputation-based consensus: A scalable and resilient mechanism for distributed mistrusting applications. *Comput. Networks*, 179:107367, 2020.

[Dou02] John R Douceur. The sybil attack. In *International workshop on peer-to-peer systems*, pages 251–260. Springer, 2002.

[Dur14] Emile Durkheim. *The division of labor in society*. Simon and Schuster, 2014.

[Dur20] Tyler Durden. Microsoft asia's a.i. 'girlfriend' has a state-imposed filter to avoid sex & politics. *LaptrinhX*, 2020.

[EFG+20a] 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.

[EFG+20b] 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.

[EGR20] El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient learning. *CoRR*, abs/2003.00010, 2020.

[EM20] El Mahdi El Mhamdi. *Robust Distributed Learning*. PhD thesis, EPFL, 2020.

[EMFGH21] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, and Lê-Nguyên Hoang. On the strategyproofness of the geometric median. *ArXiV*, 2021.

[EMGR18] 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.

[EMH21] El-Mahdi El-Mhamdi and Lê-Nguyên Hoang. On goodhart's law, with an application to value alignement. *ArXiV*, 2021.

[ESM+18] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[Fes54] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

[FG19] Brian Fung and Ahiza Garcia. Facebook has shut down 5.4 billion fake accounts this year. *CNN Business*, 2019.

[FGH21] Sadegh Farhadkhani, Rachid Guerraoui, and Lê-Nguyên Hoang. Strategyproof learning: Building trustworthy user-generated datasets. *ArXiV*, 2021.

[Fla72] Kent V Flannery. The cultural evolution of civilizations. *Annual review of ecology and systematics*, 3(1):399–426, 1972.

[FZS21] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021.

[Gac15] Cory Gackenheimer. *Introduction to React*. Apress, 2015.

[Gei16] Edward Moore Geist. It's already too late to stop the ai arms race—we must manage it instead. *Bulletin of the Atomic Scientists*, 72(5):318–321, 2016.

[Had17] 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.

[HE19] Lê Nguyên Hoang and El-Mahdi El-Mhamdi. *Le fabuleux chantier: Rendre l'intelligence artificielle robustement bénéfique*. EDP Sciences, 2019.

[HFE21] Lê Nguyên Hoang, Louis Faucon, and El-Mahdi El-Mhamdi. Recommendation algorithms, a neglected opportunity for public health. *Revue Médecine et Philosophie*, 4(2):16–24, 2021.

[HKM09] Adrian Holovaty and Jacob Kaplan-Moss. *The definitive guide to Django: Web development done right*. Apress, 2009.

[Hoa19] Lê Nguyên Hoang. Towards robust end-to-end alignment. In *SafeAI@ AAAI*, 2019.

[Hoa20] Lê Nguyên Hoang. Science communication desperately needs more aligned recommendation algorithms. *Frontiers in Communication*, 5:115, 2020.

[HP10] Sameer Hinduja and Justin W Patchin. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221, 2010.

[HS17] Nazir S Hawi and Maya Samaha. The relations among social media addiction, self-esteem, and life satisfaction in university students. *Social Science Computer Review*, 35(5):576–586, 2017.

[JKCP15] 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.

[JS61] William James and Charles Stein. Estimation with quadratic loss. In *Fourth Berkeley Symposium*, pages 361–380. University California Press, 1961.

[Lik32] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[LJMZ02] Jerry W Lee, Patricia S Jones, Yoshimitsu Mineyama, and Xinwei Esther Zhang. Cultural differences in responses to a likert scale. *Research in nursing & health*, 25(4):295–306, 2002.

[LKK+19] 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.

[May18] Lucas Maystre. *Efficient Learning from Comparisons*. PhD thesis, EPFL, 2018.

[MB12] Katherine L Milkman and Jonah Berger. What makes online content viral. *Journal of Marketing Research*, 49(2):192–205, 2012.

[MIC+16] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. Neurotics can't focus: An *in situ* study of online multitasking in the workplace. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade, editors, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 1739–1744. ACM, 2016.

[MMS+19] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[MN20] Kris McGuffie and Alex Newhouse. The radicalization risks of GPT-3 and advanced neural language models. *CoRR*, abs/2009.06807, 2020.

[MT18] Bertin Martens and Songül Tolan. Will this time be different? a review of the literature on the impact of artificial intelligence on employment, incomes and growth. *JRC Digital Economy Working Paper 2018-08*, 2018.

[Mue21] Tobias Mueller. Let's attest! multi-modal certificate exchange for the web of trust. In *International Conference on Information Networking, ICOIN 2021, Jeju Island, South Korea, January 13-16, 2021*, pages 758–763. IEEE, 2021.

[Nor10] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

[NP82] Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, pages 289–297, 1982.

[RHAV15] Stuart Russell, Sabine Hauert, Russ Altman, and Manuela Veloso. Ethics of artificial intelligence. *Nature*, 521(7553):415–416, 2015.

[Roe19] Peter Gordon Roetzel. Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business research*, 12(2):479–522, 2019.

[Rou21] Sébastien Rouault. *Practical Byzantine-resilient Stochastic Gradient Descent*. PhD thesis, EPFL, 2021.

[ROW+20] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In Mireille Hildebrandt, Carlos Castillo, Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 131–141. ACM, 2020.

[RRRR18] Hans Rosling, Anna Rosling Rönnlund, and Ola Rosling Rosling. *Factfulness: Ten Reasons We're Wrong About the World–and Why Things Are Better Than You Think*. Flatiron Books, 2018.

[Rus19] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[SAHM20] Jonathan Stray, Steven Adler, and Dylan Hadfield-Menell. What are you optimizing for? aligning recommender systems with human values. In *Participatory Approaches to Machine Learning Workshop, ICML 2020*, 2020.

[Soa15] Nate Soares. The value learning problem, 2015.

[Sol18] Joan E. Solsman. Youtube's ai is the puppet master over most of what you watch. *CNET*, 2018.

[SSP+13] Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1374–1383. The Association for Computer Linguistics, 2013.

[Ste56] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States, 1956.

[Sub16] 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.

[SW17] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017*, pages 1–10. Association for Computational Linguistics, 2017.

[Tar10] Nick Tarleton. Coherent extrapolated volition: A meta-level approach to machine ethics. *Machine Intelligence Research Institute*, 2010.

[TBB18] Ofir Turel, Damien Brevers, and Antoine Bechara. Time distortion when users at-risk for social media addiction engage in non-social media tasks. *Journal of psychiatric research*, 97:84–88, 2018.

[TCK19] Harry Chandra Tanuwidjaja, Rakyong Choi, and Kwangjo Kim. A survey on deep learning techniques for privacy-preserving. In Xiaofeng Chen, Xinyi Huang, and Jun Zhang, editors, *Machine Learning for Cyber Security - Second International Conference, ML4CS 2019, Xi'an, China, September 19-21, 2019, Proceedings*, volume 11806 of *Lecture Notes in Computer Science*, pages 29–46. Springer, 2019.

[Tig19] 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.

[VN19] Ivan Vendrov and Jeremy Nixon. Aligning recommender systems as cause area. *Effective Altruism Forum*, 2019.

[WH18] Samuel C Woolley and Philip N Howard. *Computational propaganda: political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018.

[WMCL19] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor.*, 21(2):80–90, 2019.

[Woo20] Samuel Woolley. *The reality game: how the next wave of technology will break the truth*. Hachette UK, 2020.

[WPN+19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019.

[WSM+19] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[WTL16] Fern K Willits, Gene L Theodori, and AE Luloff. Another look at likert scales. *Journal of Rural Social Sciences*, 31(3):6, 2016.

[Yud16] Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016.

[ZGLS20] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.