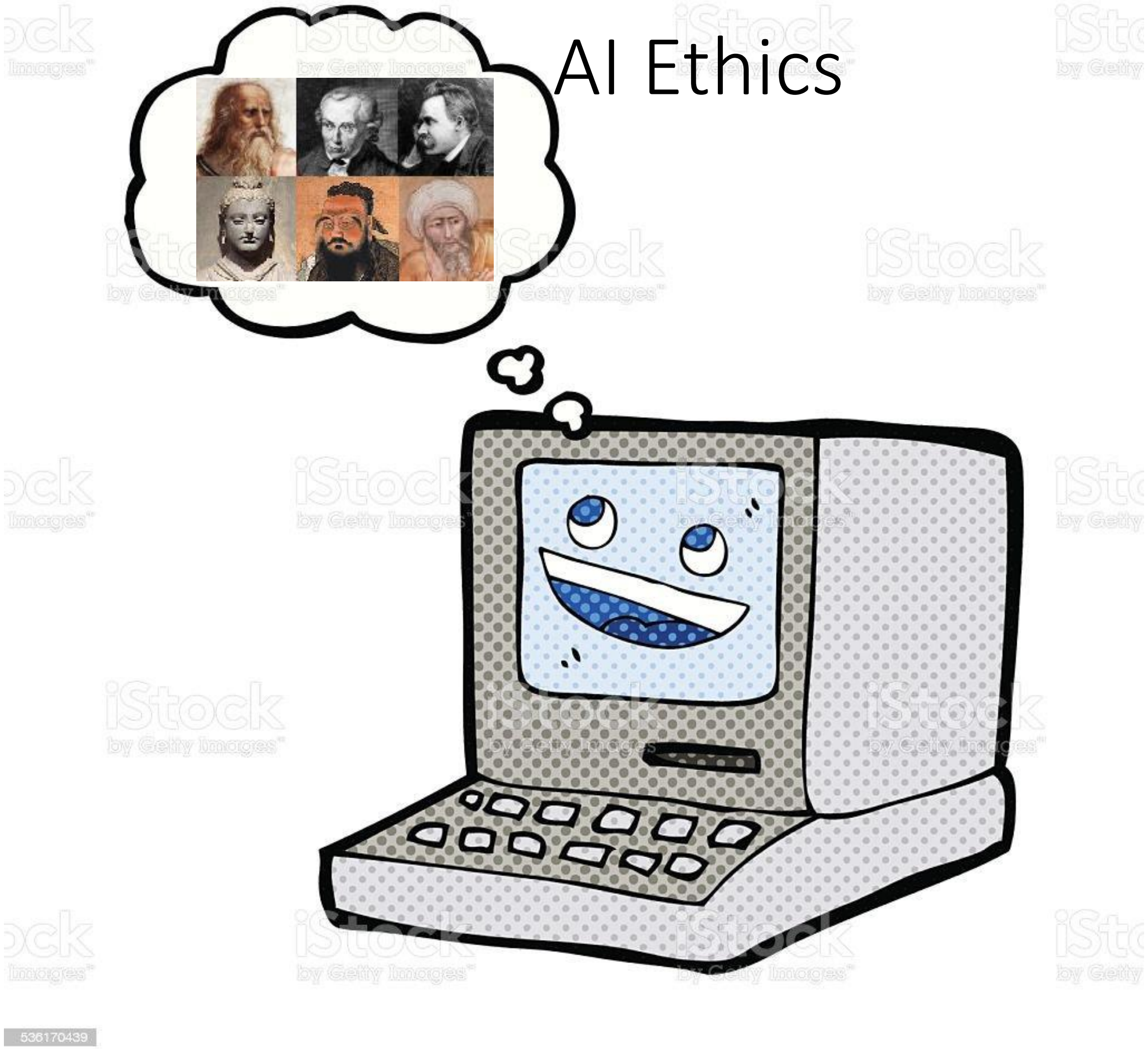
AI Ethics



ECE324, Winter 2022

Michael Guerzhoy

Ethics

- According to Wikipedia, there are three areas of ethics
  - Meta-ethics, concerning the theoretical meaning and reference of moral propositions, and how their truth values (if any) can be determined;
  - Normative ethics, concerning the practical means of determining a moral course of action;
  - Applied ethics, concerning what a person is obligated (or permitted) to do in a specific situation or a particular domain of action

AI Ethics

- Some see AI Ethics as being akin to medical ethics – just as there are guidelines about how to behave in situations involving medical patients, there ought to be guidelines about how to develop AI
    - This is more an aspiration than the current situation – there is no AI Code of Ethics
- Some see the practice of AI Ethics as simply reflecting on what one is doing, and choosing to do the right thing
    - Practicing personal ethics when developing AI

Different approaches to AI Ethics

- Decide as a community what the ethical guidelines are, and them have everyone follow the guidelines
  - Akin to medical ethics
- Encourage individuals to reflect on the ethical aspects of their work, hoping that individuals would choose to do the right thing
- Not necessarily a dichotomy

Instances of bad AI Ethics

- When they teach you medical ethics, they often point to instances where medical ethics was not followed
- We’ll take a similar approach to AI Ethics

Aside: not always a broad consensus about medical ethics

- There was a vocal group of journalists and bloggers that claimed that *not* use human challenge trials for COVID vaccines was an ethical lapse
  - In human challenge trials, healthy humans would be infected with COVID in order to quickly test vaccines/treatments
  - Human challenge trials are often considered unethical (with exceptions)
- There is a wide variety of opinion on how fetuses should be treated
- I'll give you the arguments for what are fairly widely considered instances of "Bad AI Ethics" (but in some instances are also widely used)
  - Decide for yourself!

Bad AI Ethics: COMPAS

- A system to predict recidivism, used to determine whether to grant bail or not
  - Argument: the system reproduces racially unequal outcomes
  - Argument: a computer shouldn’t decide whether to grant bail or not
  - Argument: the bail system is inherently bad and participating in it is bad
  - Counterargument: unless we get rid of the bail system, *something* needs to be used to decide whether to grant bail, and that something (an algorithm or a human) is better if it’s at least calibrated

Bad AI Ethics: ImageNet

- Argument: images collected without subjects' informed consent
  - Images could be reproduced by e.g. GANs accidentally
- Argument: images are sometimes pornographic and sometimes offensive
- Argument: labels and images encode racist and sexist attitudes
- Counterargument: to the extent that very large datasets are important for AI development, it is hard-to-impossible to curate them

Bad AI Ethics: “AI Gaydar”

- A system that was trained on photos from dating websites and predicts the sexual orientation of the person in the profile pic
- Argument: the system could be used by a government to persecute people based on their photos
- Argument: the paper implies that it’s possible to tell someone’s sexual orientation from their features, when it’s just as likely that the system is picking up on clothing/accessories/etc
    - Argument: it is bad to support the idea that physical features can be used to infer something about the person’s character

Bad AI Ethics: Inferring IQ from Essays

- Argument: bad to imply that IQ is a valid construct
- Argument: such a system is probably biased against some social groups

Bad AI Ethics: Large Language Models (LLMs) Trained on Internet-Scale Datasets

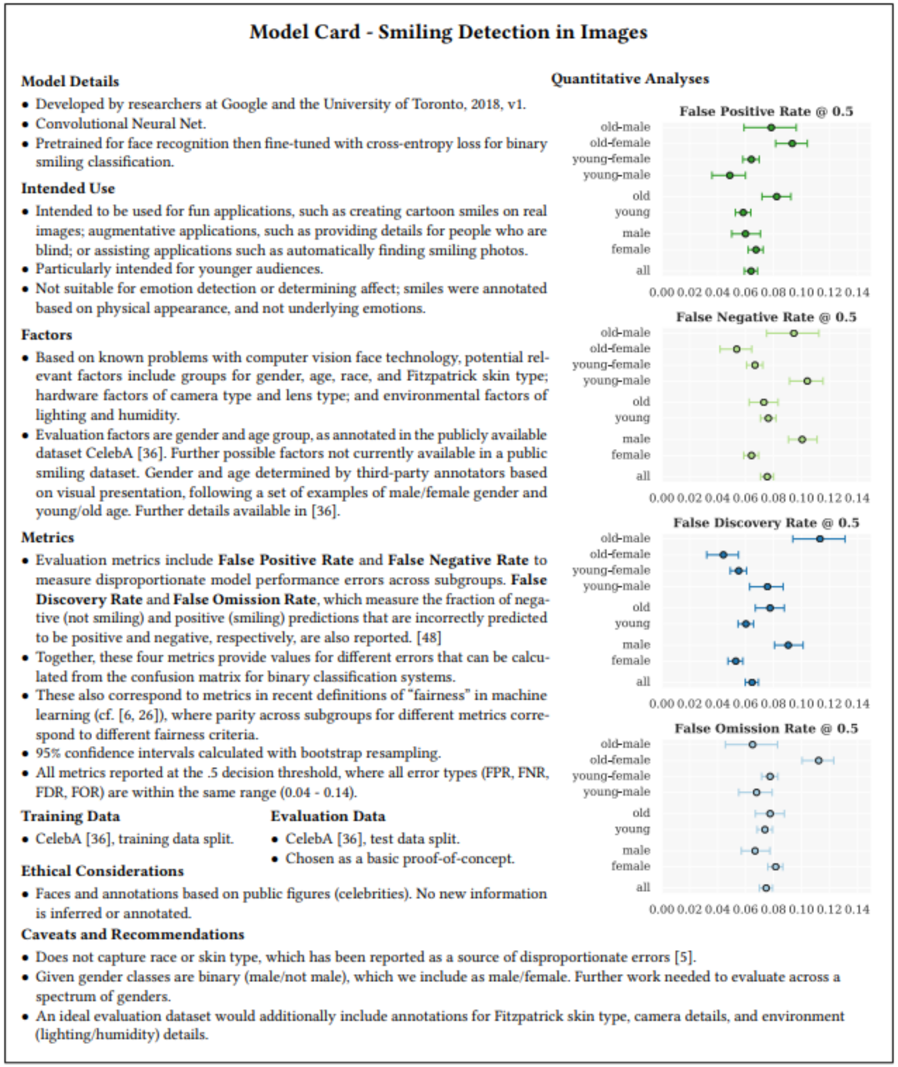
- Argument: training LLMs is expensive
  - Benefits accrue to rich companies that train LLMs and speakers of English (and a few other languages)
  - Costs from effects CO2 emissions are borne by people in countries that are often not English-speaking (Counterargument: the CO2 emissions are not that large)
- Argument: training on internet-scale data necessarily means that the language of internet users (disproportionately rich and white) is prioritized in modelling
- Argument: training on internet-scale data means attitudes and biases existing on the internet, such as racism and sexism, are encoded in the model
- Argument: prioritizing research on LLMs takes time and resources away from other approaches that might work on “low-resource languages” (languages for which internet-scale data is not available)
- Argument: LLMs aren’t the right avenue to pursue to develop true AI anyway, so training them wastes resources

Making AI practitioners more ethical

- Encourage individuals to reflect on the ethical aspects of their work, hoping that individuals would choose to do the right thing

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

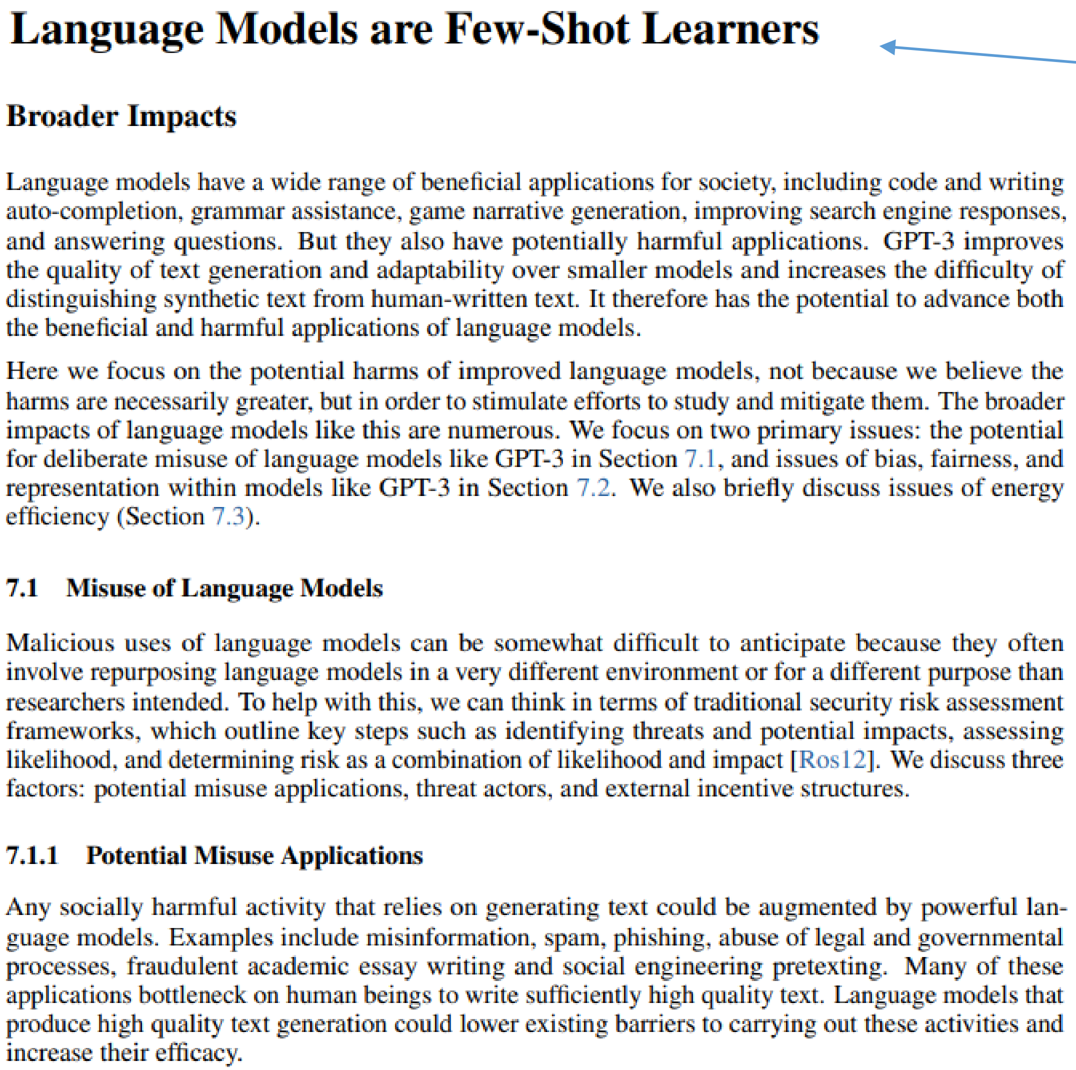
Quantitative Analyses

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Making AI practitioners more ethical

- One of the premier machine learning conferences, NeurIPS, required a “broader impact” statement from all papers in 2020
  - Made more optional in 2021



Language Models are Few-Shot Learners

Broader Impacts

Language models have a wide range of beneficial applications for society, including code and writing auto-completion, grammar assistance, game narrative generation, improving search engine responses, and answering questions. But they also have potentially harmful applications. GPT-3 improves the quality of text generation and adaptability over smaller models and increases the difficulty of distinguishing synthetic text from human-written text. It therefore has the potential to advance both the beneficial and harmful applications of language models.

Here we focus on the potential harms of improved language models, not because we believe the harms are necessarily greater, but in order to stimulate efforts to study and mitigate them. The broader impacts of language models like this are numerous. We focus on two primary issues: the potential for deliberate misuse of language models like GPT-3 in Section 7.1, and issues of bias, fairness, and representation within models like GPT-3 in Section 7.2. We also briefly discuss issues of energy efficiency (Section 7.3).

7.1 Misuse of Language Models

Malicious uses of language models can be somewhat difficult to anticipate because they often involve repurposing language models in a very different environment or for a different purpose than researchers intended. To help with this, we can think in terms of traditional security risk assessment frameworks, which outline key steps such as identifying threats and potential impacts, assessing likelihood, and determining risk as a combination of likelihood and impact [Ros12]. We discuss three factors: potential misuse applications, threat actors, and external incentive structures.

7.1.1 Potential Misuse Applications

Any socially harmful activity that relies on generating text could be augmented by powerful language models. Examples include misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting. Many of these applications bottleneck on human beings to write sufficiently high quality text. Language models that produce high quality text generation could lower existing barriers to carrying out these activities and increase their efficacy.

GPT-3 paper

7.1.2 Threat Actor Analysis

Threat actors can be organized by skill and resource levels, ranging from low or moderately skilled and resourced actors who may be able to build a malicious product to 'advanced persistent threats' (APTs): highly skilled and well-resourced (e.g. state-sponsored) groups with long-term agendas [SBC+19].

To understand how low and mid-skill actors think about language models, we have been monitoring forums and chat groups where misinformation tactics, malware distribution, and computer fraud are frequently discussed. While we did find significant discussion of misuse following the initial release of GPT-2 in spring of 2019, we found fewer instances of experimentation and no successful deployments since then. Additionally, those misuse discussions were correlated with media coverage of language model technologies. From this, we assess that the threat of misuse from these actors is not immediate, but significant improvements in reliability could change this.

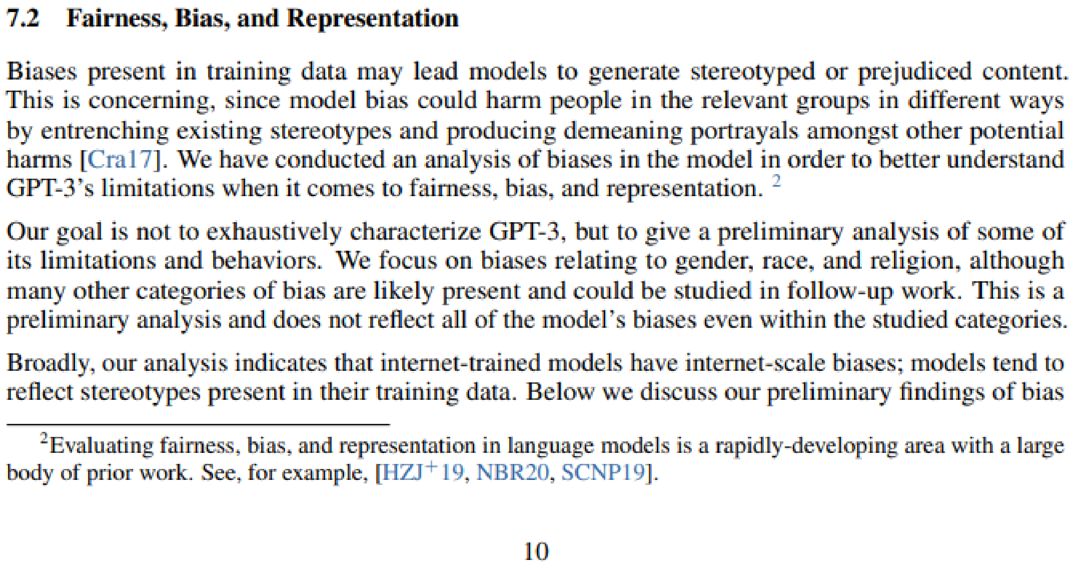
Because APTs do not typically discuss operations in the open, we have consulted with professional threat analysts about possible APT activity involving the use of language models. Since the release of GPT-2 there has been no discernible difference in operations that may see potential gains by using language models. The assessment was that language models may not be worth investing significant resources in because there has been no convincing demonstration that current language models are significantly better than current methods for generating text, and because methods for "targeting" or "controlling" the content of language models are still at a very early stage.

7.1.3 External Incentive Structures

Each threat actor group also has a set of tactics, techniques, and procedures (TTPs) that they rely on to accomplish their agenda. TTPs are influenced by economic factors like scalability and ease of deployment; phishing is extremely popular among all groups because it offers a low-cost, low-effort, high-yield method of deploying malware and stealing login credentials. Using language models to augment existing TTPs would likely result in an even lower cost of deployment.

Ease of use is another significant incentive. Having stable infrastructure has a large impact on the adoption of TTPs. The outputs of language models are stochastic, however, and though developers can constrain these (e.g. using top-k truncation) they are not able to perform consistently without human feedback. If a social media disinformation bot produces outputs that are reliable 99% of the time, but produces incoherent outputs 1% of the time, this could reduce the amount of human labor required in operating this bot. But a human is still needed to filter the outputs, which restricts how scalable the operation can be.

Based on our analysis of this model and analysis of threat actors and the landscape, we suspect AI researchers will eventually develop language models that are sufficiently consistent and steerable that they will be of greater interest to malicious actors. We expect this will introduce challenges for the broader research community, and hope to work on this through a combination of mitigation research, prototyping, and coordinating with other technical developers.

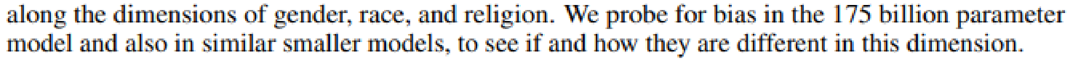


7.2 Fairness, Bias, and Representation

Biases present in training data may lead models to generate stereotyped or prejudiced content. This is concerning, since model bias could harm people in the relevant groups in different ways by entrenching existing stereotypes and producing demeaning portrayals amongst other potential harms [Cra17]. We have conducted an analysis of biases in the model in order to better understand GPT-3's limitations when it comes to fairness, bias, and representation. [2]

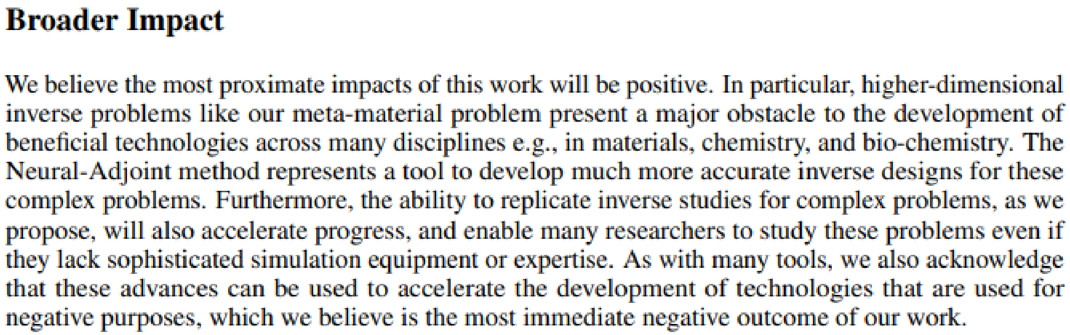
Our goal is not to exhaustively characterize GPT-3, but to give a preliminary analysis of some of its limitations and behaviors. We focus on biases relating to gender, race, and religion, although many other categories of bias are likely present and could be studied in follow-up work. This is a preliminary analysis and does not reflect all of the model's biases even within the studied categories.

Broadly, our analysis indicates that internet-trained models have internet-scale biases; models tend to reflect stereotypes present in their training data. Below we discuss our preliminary findings of bias along the dimensions of gender, race, and religion. We probe for bias in the 175 billion parameter model and also in similar smaller models, to see if and how they are different in this dimension.



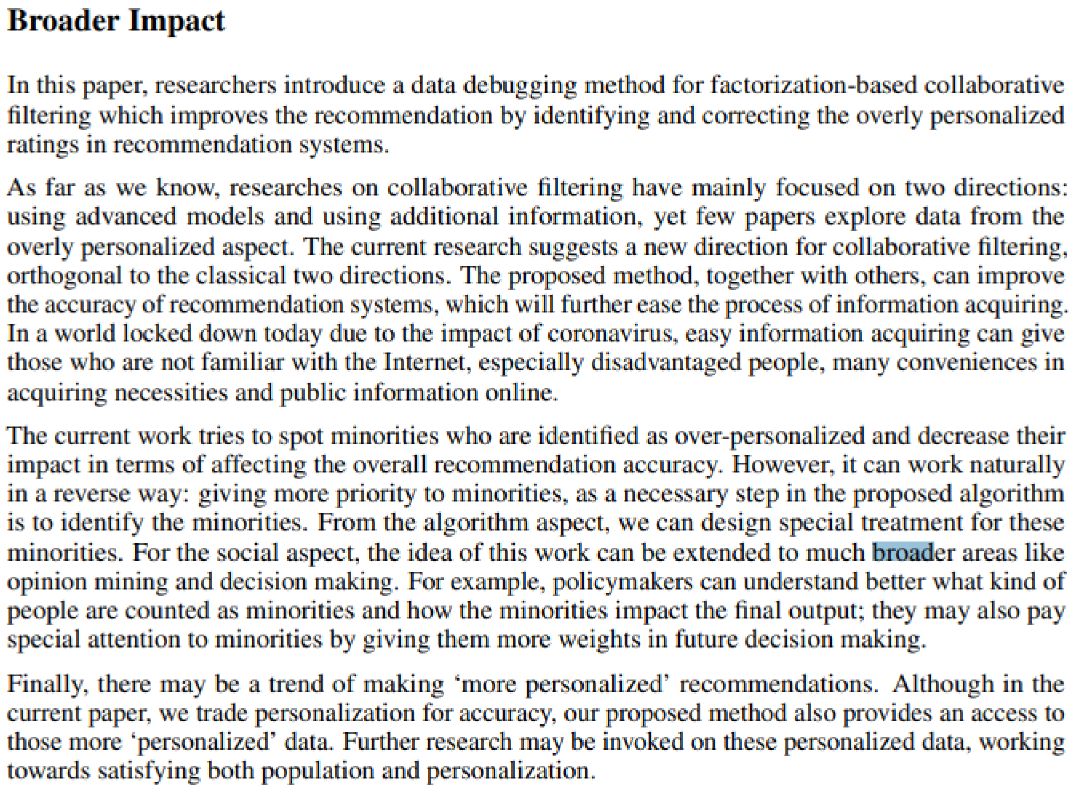
[2] Evaluating fairness, bias, and representation in language models is a rapidly-developing area with a large body of prior work. See, for example, [HZJ+19, NBR20, SCNP19].

Other NeuRIPS papers



Broader Impact

We believe the most proximate impacts of this work will be positive. In particular, higher-dimensional inverse problems like our meta-material problem present a major obstacle to the development of beneficial technologies across many disciplines e.g., in materials, chemistry, and bio-chemistry. The Neural-Adjoint method represents a tool to develop much more accurate inverse designs for these complex problems. Furthermore, the ability to replicate inverse studies for complex problems, as we propose, will also accelerate progress, and enable many researchers to study these problems even if they lack sophisticated simulation equipment or expertise. As with many tools, we also acknowledge that these advances can be used to accelerate the development of technologies that are used for negative purposes, which we believe is the most immediate negative outcome of our work.



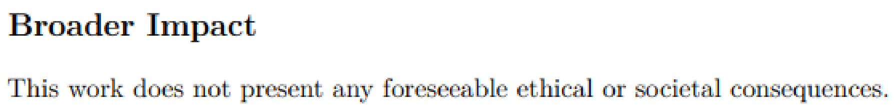
Broader Impact

In this paper, researchers introduce a data debugging method for factorization-based collaborative filtering which improves the recommendation by identifying and correcting the overly personalized ratings in recommendation systems.

As far as we know, researches on collaborative filtering have mainly focused on two directions: using advanced models and using additional information, yet few papers explore data from the overly personalized aspect. The current research suggests a new direction for collaborative filtering, orthogonal to the classical two directions. The proposed method, together with others, can improve the accuracy of recommendation systems, which will further ease the process of information acquiring. In a world locked down today due to the impact of coronavirus, easy information acquiring can give those who are not familiar with the Internet, especially disadvantaged people, many conveniences in acquiring necessities and public information online.

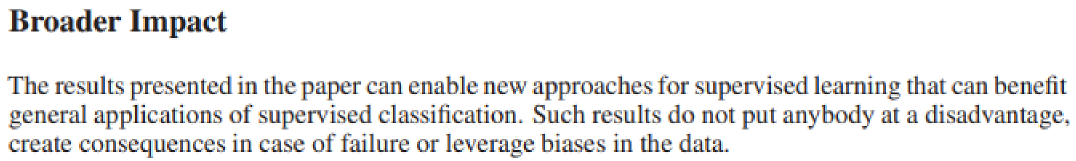
The current work tries to spot minorities who are identified as over-personalized and decrease their impact in terms of affecting the overall recommendation accuracy. However, it can work naturally in a reverse way: giving more priority to minorities, as a necessary step in the proposed algorithm is to identify the minorities. From the algorithm aspect, we can design special treatment for these minorities. For the social aspect, the idea of this work can be extended to much broader areas like opinion mining and decision making. For example, policymakers can understand better what kind of people are counted as minorities and how the minorities impact the final output; they may also pay special attention to minorities by giving them more weights in future decision making.

Finally, there may be a trend of making 'more personalized' recommendations. Although in the current paper, we trade personalization for accuracy, our proposed method also provides an access to those more 'personalized' data. Further research may be invoked on these personalized data, working towards satisfying both population and personalization.



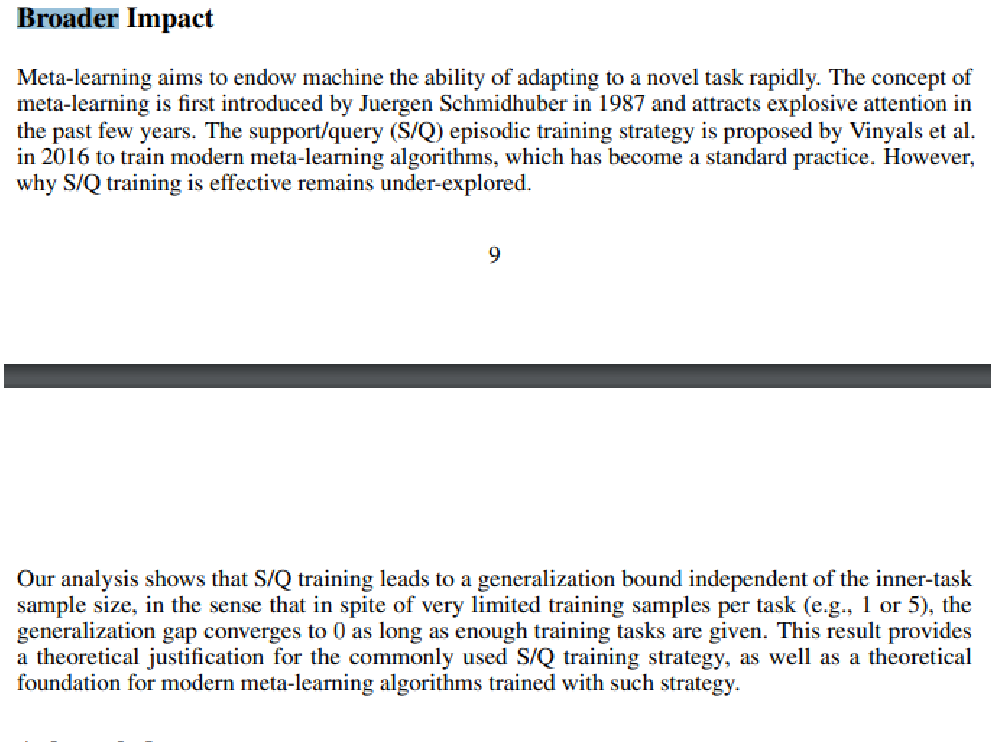
Broader Impact

This work does not present any foreseeable ethical or societal consequences.



Broader Impact

The results presented in the paper can enable new approaches for supervised learning that can benefit general applications of supervised classification. Such results do not put anybody at a disadvantage, create consequences in case of failure or leverage biases in the data.



Broader Impact

Meta-learning aims to endow machine the ability of adapting to a novel task rapidly. The concept of meta-learning is first introduced by Juergen Schmidhuber in 1987 and attracts explosive attention in the past few years. The support/query (S/Q) episodic training strategy is proposed by Vinyals et al. in 2016 to train modern meta-learning algorithms, which has become a standard practice. However, why S/Q training is effective remains under-explored.

9

Our analysis shows that S/Q training leads to a generalization bound independent of the inner-task sample size, in the sense that in spite of very limited training samples per task (e.g., 1 or 5), the generalization gap converges to 0 as long as enough training tasks are given. This result provides a theoretical justification for the commonly used S/Q training strategy, as well as a theoretical foundation for modern meta-learning algorithms trained with such strategy.

Your broader impact statement

- Think about the consequences for a variety of stakeholders
- Think about the uses and misuses of the technology you build
- Sometimes the honest answer is “I don’t know”
    - In the project, explain *why* that is the answer

For and against broader impact statements

- Argument: a way to “nudge” researchers and practitioners toward thinking about the ethical implications of their work
    - Everyone is responsible to not make the world a worse place
- Counterargument: a requirement necessarily produces shallow and empty statements
    - Countercounterargument: well maybe papers with shallow and empty statements should be rejected
    - Countercountercounterargument: but often shallow and empty is genuinely the best anyone can do