OpenAI Report
November, 2019

Release Strategies and the Social Impacts of Language Models

Irene Solaiman*
OpenAI
irene@openai.com

Miles Brundage
OpenAI
miles@openai.com

Jack Clark
OpenAI
jack@openai.com

Amanda Askell
OpenAI
amanda@openai.com

Ariel Herbert-Voss
Harvard University
ariel_herbertvoss@g.harvard.edu

Jeff Wu
OpenAI
jeffwu@openai.com

Alec Radford
OpenAI
alec@openai.com

Gretchen Krueger
OpenAI
gretchen@openai.com

Jong Wook Kim
OpenAI
jongwook@openai.com

Sarah Kreps
Cornell University
sarah.kreps@cornell.edu

Miles McCain
Politiwatch
miles@rmrm.io

Alex Newhouse
CTEC
anewhouse@middlebury.edu

Jason Blazakis
CTEC
jblazakis@middlebury.edu

Kris McGuffie
CTEC
Kmcguffie@middlebury.edu

Jasmine Wang
OpenAI
jasmine@openai.com

*Listed in descending order of contribution.

Contents

Overview

GPT-2 is a large-scale unsupervised language model that generates coherent paragraphs of text, first announced by OpenAI in February 2019 [65]. We developed four variants of the model, ranging in size from small (124 million parameters) to large (~1.5 billion parameters). We chose a staged release process, releasing the smallest model in February, but withholding larger models due to concerns about the potential for misuse, such as generating fake news content, impersonating others in email, or automating abusive social media content production [56]. We released the 355 million parameter model in May as part of a staged release process. We released our 774 million parameter model in August with a six-month follow up announcement, and we are now releasing our 1.5 billion parameter model.

While large language models' flexibility and generative capabilities raise misuse concerns, they also have a range of beneficial uses - they can assist in prose, poetry, and programming; analyze dataset biases; and more. We want to release systems that will have a widely-distributed positive impact on society and have low misuse potential, and have striven to make release decisions informed by analysis, engagement, and empirical evidence.

Instead of releasing the full 1.5 billion model in February, we adopted a 'staged release' process. This delay of nine months allowed time between model releases to conduct risk and benefit analyses as model sizes increased. We also hope our staged release process was helpful in allowing others time to adapt and react: giving researchers a chance to mitigate risk of potential misuse, and giving the general public time to adapt to a world in which it is prudent to mistrust everything they read a little more. In addition to finding minimal evidence of misuse so far, several other factors contributed to our confidence in publishing our 774 million and 1.5 billion parameter models. These include what we learned about the positive social impact of beneficial uses, and what we learned through our partnerships among the AI community and through discussions across fields about establishing norms for responsible publication. This report discusses OpenAI's work related to staged release of large models, partnership-based research, and broader issues in responsible publication that the AI community will need to address.

1 Staged Release

In February 2019, we released the 124 million parameter GPT-2 language model. In May 2019, we released the 355 million parameter model and a dataset of outputs from all four models (124 million, 355 million, 774 million, and 1.5 billion parameters) to aid in training humans and classifiers to detect synthetic text, and assessing biases encoded in GPT-2 generated outputs. In August, we released our 774 million parameter model along with the first version of this report and additional release documentation on GitHub. We are now releasing our 1.5 billion parameter version of GPT-2 with this updated report and updated documentation.

As performance across dimensions - such as the reliability of generating coherent text - tends to improve with model size, we decided not to release all four GPT-2 models simultaneously due to concerns about the larger models being misused. By staggering releases as part of staged release, we allow time for risk analyses and use findings from smaller models to inform the actions taken with larger ones.

Since February 2019, we have communicated with researchers who created similar language models to GPT-2. We have also seen other labs approach their own language model research with a similarly cautious mindset to the staged release; for example, Allen Institute for Artificial Intelligence and University of Washington researchers adopted an incremental approach when releasing their GROVER model [81]. GROVER researchers also performed in-depth threat modeling and discussed their findings with other AI researchers, including those at OpenAI. Similarly, NLP company Hugging Face decided not to release some of its internal language models and provided educational information about the limitations of chatbots alongside its latest release [19]. Finally, AI company AI21 recently announced work on controllable neural text generation, and noted that their demo was based on a model equivalent in size to public versions of GPT-2 and GROVER [42]. Students working independently at the Technical University of Munich and Brown University replicated GPT-2 and wrote about their respective views on responsible publication, with one choosing not to publish[2] and another group publishing a similar model to GPT-2 (in part to demonstrate the feasibility of doing so) [28]. Finally, Salesforce released their more controllable large language model, CTRL, [39] along with an analysis of the societal implications of pretrained models [73].

should this be standard? Always release model w/ risk analysis?

To accompany our staged release process, we formed partnerships, held discussions with researchers, observed GPT-2 uses, and conducted in-house research into automated detection, biases, and misuse potential. We remain cautiously optimistic about the social benefit of our larger language models.

[2] Connor Leahy at the Technical University of Munich wrote about his intent to publish a replicated version of GPT-2 but changed his mind after discussion with researchers [43] [44].

2 Partnerships

We established partnerships with four leading organizations that are studying potential malicious uses of GPT-2, examining how to detect GPT-2-generated text, analyzing how humans respond to text generated by GPT-2, and studying biases in GPT-2 outputs.

When forming partnerships, we signed a non-commercial legal agreement with a partner organization to provide our model for their research use, and/or we provided a partner organization with a secure sampling interface to the larger models. This involved extensive negotiation with prospective partners to reach an agreement that satisfied all parties.[3] We believe similar partnerships will be increasingly important as AI systems become more powerful and are publishing a generic version of the legal agreement we developed [see Appendix A].

We are excited to be partnering with the following organizations to study GPT-2:

- **Cornell University** is studying human susceptibility to digital disinformation generated by language models.
- **The Middlebury Institute of International Studies** Center on Terrorism, Extremism, and Counterterrorism (CTEC) is exploring how GPT-2 could be misused by terrorists and extremists online.
- **The University of Oregon** is developing a series of "bias probes" to analyze bias within GPT-2.
- **The University of Texas at Austin** is studying the statistical detectability of GPT-2 outputs after fine-tuning the model on domain-specific datasets, as well as the extent of detection transfer across different language models.

Our partners at Middlebury's CTEC gave us insights not only on misuse capabilities, but also on detection countermeasures [see Appendix D]. Our partners at Cornell University highlighted the diminishing returns to larger models from a human detection perspective [see Appendix E]. Ongoing partner research brings new perspectives to misuse, detection, and bias analysis and contributes to evidence for informing release decisions. Our hope is that partnerships can be a scalable tool for studying and mitigating downsides of powerful models, in order to enable us to unlock benefits in a responsible manner.

ie more researchers working together will help ID more risks

[3] We are grateful to all prospective partners who took the time to discuss these issues with us, regardless of whether we ended up partnering.

3 Engagement

In addition to the partnerships above, we have been contributing to the Partnership on AI (PAI)'s ongoing work on developing responsible publication norms for machine learning and AI, and co-hosted a discussion on the topic to source input from across the AI ecosystem.[4] Our work with PAI explores possible mechanisms to maximize the benefits of open publication while mitigating the risks of advanced ML systems via approaches such as staged release and internal review processes.[5] By sharing the insights learned from our experience releasing GPT-2, we hope to contribute to the continued efforts of the community to navigate these issues.

We also discussed impacts of GPT-2 and large language models with members of the AI community, researchers, companies potentially targeted by disinformation campaigns, and activists who work on topics like digital disinformation and online abuse. We also spoke about GPT-2 and our approach to releasing it at a speech at the AI for Social Good workshop at ICLR and a range of other venues, including Congress.[6]

[4] PAI is keen to engage with a broad range of stakeholders in the AI/ML community on this project. If you would like to participate, please contact rosie@partnershiponai.org.

[5] Although the project is in its early phases, a number of PAI Partner organizations are already trialling processes built upon it. This includes Salesforce's decision to publish CTRL, and Facebook, Microsoft, and Amazon's use of a PAI steering committee to inform the design of their Deepfake Detection Challenge.

[6] This includes a Scaled Machine Learning Conference talk from Ilya Sutskever [70], a guest lecture by Alec Radford at UC Berkeley [64], a TWIML podcast including Miles Brundage and Amanda Askell [37], and a US Global Engagement Center talk by Jack Clark.

4 Social Impacts of Large Language Models

Large language models have a wide range of usages across domains. Some uses include:

- Generating text from the model "out of the box" (e.g. zero-shot generation);
- Generating specific styles of text after the model has been trained further (fine-tuned) on a different dataset;
- Creating task-specific systems (e.g. sentiment classifiers, speech recognition systems, translation systems, dialogue systems), often with less data and computing power than would be needed to build systems from scratch;
- Discriminating between synthetic text generated by a language model (especially adversarial examples) and human-authored text; and
- Analyzing model activations and outputs scientifically to understand its knowledge and biases.

4.1 Beneficial Use Potential

There are many active beneficial applications of language models. These include biomedical literature analysis [7], generating synthetic test data [31], and generating radiology reports [46] and EEG reports [10]. Other language models have accelerated NLP research and applications by providing better starting points for supervised training models [17], introducing techniques for fine-tuning [36], and enhancing performance in challenges like question answering and sentiment analysis [63]. These techniques help researchers, practitioners, and users.

We have seen GPT-2 in particular used in the domains listed below:

Domain	Use
Software Engineering	Code Autocompletion [71]
Writing	Grammar Assistance [3] Autocompletion-Assisted Writing [20]
Art	Creating or Aiding Literary Art [69; 74; 24] Poetry Generation [11]
Entertainment	Gaming [75] Chatbots [77; 55; 12]
Health	Medical Question-Answering systems[7] [32]

[7] Note that in a safety-critical domain such as medicine, understanding the biases encoded in AI systems is especially important, and as such the author emphasizes that Doc Product is intended as a proof of concept rather than a production system.

The diversity of GPT-2's early applications gives us confidence that releasing larger model sizes will enable further benefits. A prominent GPT-2 application is in aiding the writing process, both in natural and programming languages. Grammarly published a paper highlighting GPT-2's utility in grammatical error correction [3]. Hugging Face developed a web-based writing UI with a document editor-like interface, where writers can iteratively generate text [20]. Deep TabNine is an all-language auto-completion tool trained on approximately two million GitHub files that intends to enhance software developers' workflows [71].[8]

With more fine-grained control over outputs, generative models could be better applied across domains. In OpenAI's MuseNet, a generative model of music, creators can directly interact with the generative model in the advanced mode to specify instruments and composers and influence the distribution of the model's suggestions [61]. GPT-2 Explorer, developed by the Allen Institute for Artificial Intelligence, displays the probabilities that GPT-2 assigns to various possible next words in a sequence [25]. It provides a separate, autocomplete-like interface to better understand GPT-2's capabilities and limitations. Further improvements on models and interfaces will likely yield further scientific, creative, and commercial applications.

→ well, if you don't do the research someone else will & they might not be as good at two shoes about ethics!

4.2 Misuse: Actor Assessment

In our initial post on GPT-2, we noted our concern that its capabilities could lower costs of disinformation campaigns, although we were unsure about how to best characterize such risks. We have since further researched the digital disinformation landscape, the feasibility of disinformation-related misuse cases, and other potential misuses of language models. We drew on external engagement with security experts and the AI community, monitoring of websites and anonymous forums with a history of spreading disinformation and organizing hate movements, discussions with policymakers in defense and intelligence, and proofs of concept to inform our staged release decisions.

We have broken down malicious actors into three tiers, organized in ascending order by increasing levels of skill and resources:

1. Low-skilled, limited resource actors who may be ideologically motivated or simply curious in their abilities. They may attempt to alter training data to bias a language model.
2. Actors with moderate programming skills and resources who are able and willing to build a malicious product, such as tools for webspam.
3. Advanced persistent threats (APTs): highly skilled and well-resourced groups, like state-sponsored actors, that have a long-term agenda.

[8] Disclosure: Deep TabNine was developed by a former OpenAI intern.

At all tiers, malicious actors could be motivated by the pursuit of monetary gain, a particular political agenda, and/or a desire to create chaos or confusion. The thought processes and machinations of the two lower-tiered of actors are often easier to observe. We have closely monitored online communities for evidence of interest in weaponizing language models; such public forums are often used to coordinate online disinformation or abuse campaigns. APT actions are notoriously difficult to monitor and mitigate.

(1) Low-skilled actors tend to interact with AI systems in an unsophisticated way, but this can still lead to harmful outcomes. A canonical example is Microsoft's "Tay" chatbot, a Twitter bot that replied based on interactions with Twitter users. Internet trolls Tweeted intentionally offensive phrases at Tay, effectively poisoning its dataset and exploiting its API, resulting in offensive Tweets. Microsoft removed the bot and released an apology that included a commitment to think more carefully about potential misuses [45]. Since GPT-2 is a trained model and not a complete interface, dataset poisoning is unlikely, but GPT-2 is at higher risk of malicious prompts and context forcing. Future products will need to be designed with malicious interaction in mind.

(2) Actors with moderate programming skills and resources have the capabilities to build tools to interface with GPT-2. Malicious uses developed by these actors could include generating fake news articles or building spambots for forums and social media. Since the initial release, Reddit and Discord bot interfaces have been built for GPT-2 and shared via popular open source channels. While there are positive uses for these tools, the potential for malicious use is high given that many malicious groups use those discussion forums to organize. However, integrating these tools into an ecosystem is a slow process and our analyses indicate minimal immediate risk of a fully-integrated malicious application using these or other interfaces developed by mid-range actors.

(3) Advanced persistent threats (APTs) are most likely to have the resources and motivation to misuse GPT-2, but APT motivations and behaviors are harder to analyze and observe, even with expert input. Governments and companies that specialize in tools and services for tracking APTs are better equipped to handle this level of threat actor. Given the specialization required, OpenAI cannot devote significant resources to fighting APT actors. OpenAI does, however, support initiatives and help develop strategies to defend against APTs enabled by GPT-2 through partnerships with external research groups. This is seen in our work with the Middlebury Institute's Center on Terrorism, Extremism, and Counterterrorism (CTEC) and Cornell University, as well as participation in conferences and workshops on related topics.

actively creating mis/dis/malinformation

Our threat monitoring did not find evidence of GPT-2 direct misuse in publicly-accessible forums but we did see evidence of discussion of misuse. Discussions had declined by our mid-May release. In cases where online actors discussed misusing GPT-2, the actors also demonstrated limited technical understanding of ML, suggesting a low likelihood of carrying out non-trivial attacks. We believe discussion among these actors was due to media attention following GPT-2's initial release; during follow-up mon-

itoring there was no indication that these actors had the resources, capabilities, or plans to execute at this time. We also found no clear malicious code sharing or large-scale misuse, and only a small number of cases of explicit public plans for misuse. This does not preclude future visible misuses, and proactive monitoring and modeling of the threat landscape will be necessary going forward. It also does not rule-out misuse, as certain actors - like those at nation-state scale - are more difficult to monitor and analyze. We are also aware that several governments have experimented with GPT-2 and other language models.

1.5 Billion Parameter Model: Threat Landscape

While the landscape for possible misuse has changed since the time of our initial release, we have not seen any significant action toward misuse language models during this time. Our current threat analysis methodology involves monitoring public discussion spaces as early indicators of private development. We have seen some discussion around GPT-2's potential to augment high-volume/low-yield operations like spam and phishing. However, we have not seen any progress (evidence of writing code or documentation) toward realizing this beyond discussion. This does not mean that difficult-to-observe high-skill threat actors like sophisticated criminal groups or nation states are not conducting work in this area, but it does indicate that threats from lower-tier threat actors are not as immediate as we previously thought. Tweaking language model outputs to consistently generate convincing template messages without significant human oversight is still difficult. However, this incentivizes the eventual creation of a public-facing API for producing synthetic text at scale. Some parallels can be drawn between this situation and the DeepFakes App or the LOIC DDoS tool, in that easy-to-use interfaces can enable malicious use from otherwise unskilled actors. This is a substantial threat but hard to predict exactly when it might occur. We will continue to monitor the situation and increase the capacity for other stakeholders in the ecosystem to assist with misuse detection and mitigation.

ie call in the experts so we can WK together

Since we have already described and released the smaller GPT-2 model, "security through obscurity" is not a valid release strategy going forward because motivated actors can still replicate results even if we choose not to release. Therefore, encountering examples of misuse in the wild will affect the timing of our release decisions and will require us to alert affected stakeholders and coordinate to determine a plan of action. Given the scale of AI's potential effects, we think it remains an open question as to what the appropriate heuristics are for such notification procedures, and it will require close collaboration between AI researchers, security professionals, potentially affected stakeholders, and policymakers, to determine appropriate approaches.

Our Partner's Work

The Middlebury's CTEC has been exploring how GPT-2 could be misused by terrorists and extremists online. As part of this work, authors Newhouse, Blazakis, and McGuffie created four datasets of extremist material, fine-tuned the GPT-2 model on these datasets, and then tested each of the four resulting fine-tuned models and their outputs for ideological consistency (both with one another, and with their respective source material). Given imprecision and other challenges associated with devising an 'ideology score,' they measured proxies for ideology. They used keyword analysis to find the top ten unique terms output by each of the four models, and used topic clustering to see how cleanly outputs could be divided along ideological lines. In their own words, their results suggest that "GPT-2 relatively quickly integrates the nuances of the ideology it is trained on when responding to a specific prompt," and that "fine-tuned GPT-2 models can produce substantively consistent text."

Results from CTEC's initial work assessing current detection methods indicate that fine-tuning significantly reduces the zero-shot detection capability of the GROVER model [81]. Despite low accuracy in labeling content generated using fine-tuned models as "fake", GROVER does manage to correctly label a small percent of the generated texts as fake without dipping below near-100% accuracy in labeling "real" human-generated text as such. This means that even if only one or two percent of the outputs from a specific network or actor are labeled fake, one can have reasonable suspicion that a neural language model is in use.

In addition to this initial work, CTEC has plans to broaden their quantitative approach, to conduct an "in-depth qualitative linguistic analysis" on model outputs, and to run "a survey to observe the abilities for both extremism experts and non-experts to distinguish between real and fake extremist texts". [See Appendix D for further results]

4.3 Detecting Synthetic Text

One key variable affecting the social impact of language models is the extent to which humans and machines can detect outputs. We found reasons for optimism as well as reasons to continue being vigilant about the misuse of language models going forward. Our thoughts on detection at this time are:

- Humans can be deceived by text generated by GPT-2 and other successful language models, and human detectability will likely become increasingly more difficult.
- Humans can improve their ability to identify synthetic text by leveraging visualization tools [27].
- Methods for statistical detection and generation are varied and may evolve further in a cat and mouse game. For example, we might use better ML systems to improve detection accuracy, but the adversary might then use better systems for generation. The adversary can also choose a dataset for fine-tuning, different sampling techniques (rejection sampling, nucleus sampling, etc), and more.
- Metadata will continue to be central to combating malicious activity online, regardless of language model output detectability. In the limit of generation capabilities, content-based detection methods would be insufficient, as generations would mimic the true distribution of human text.

A combination of human education on language models' limitations, improved model documentation, easily available tools for fine-grained analysis, and metadata-oriented approaches will improve detection capabilities. Furthermore, Schuster et al. [67] note the challenges that legitimate uses of language models raise for addressing language model misuse via detection.

We discuss our and others' research on these topics below.

Human Detection

Over the past six months, we have seen substantial research into the ability of humans to discriminate between human- and machine-generated text samples.

Research on human perception of generated text suggests that the quality of outputs increases with model size at least up until the 774 million parameter model. With a human-in-the-loop, GPT-2 can generate outputs that humans find credible. Kreps and McCain at Cornell University found that cherry-picked fake news samples from the 355 million parameter version of GPT-2 were considered "credible" about 66% of the time.[9] Similarly cherry-picked outputs from the 774 million and 1.5 billion parameter versions of

[9] GPT-2 was used to generate continuations of a real New York Times article using the first one or two paragraphs as a prompt. Each of the three model sizes (355M, 774M, and 1.5B) was used to generate 20 outputs, and the most readable 3 or 4 were selected from each set of 20 outputs.

GPT-2 were rated statistically similarly to real New York Times articles at around 75%, although output quality was mixed even among these cherry-picked samples. For example, one 774 million parameter generation received a higher score than the real article or the 1.5 billion parameter outputs. These results suggest that improved interfaces or improved sampling methods, such as nucleus sampling, could make GPT-2 more effective at generating seemingly credible text.

Kreps and McCain did a follow-up study in which they extended these results to better understand the difference in misuseability across model sizes. First, they used a fully-automated text generation pipeline,[10] removing the need for human cherry-picking and more closely resembling some of the real world use cases that we are concerned about (e.g. large-scale spam/disinformation). Second, the authors tested more of GPT-2's outputs, giving richer insight into the distribution of output qualities as opposed to just the models' peak generation ability.[11] Third, they investigated the underlying factors driving people's credibility perceptions. The authors developed a credibility score composed of independent clarity, accuracy, and believability scores. By breaking credibility down into parts and also soliciting free-form responses from survey participants, the authors identified many instances of participants explaining away inaccuracies in GPT-2 outputs. Participants who noted inaccuracies or lack of in-text sources still cited the story's plausibility as their basis for their assigned credibility score. content vs content ∴

These results help explain why there is not an even larger gap in credibility scores between model sizes: believability and clarity vary less across model sizes than accuracy does, and believability is more important than accuracy, as people often tend to explain away inaccuracies. These results give further reason to invest in educating the public about the potential misuses of language models, since the results suggest high credulity among respondents. Finally by analyzing new data across model sizes, the authors found that the difference between the 774 million parameter model and the 1.5 billion parameter model is smaller than that between 355 million and 774 million parameter models, and relates primarily to greater peak performance rather than greater mean performance.[12] [See Appendix E for further results]

Finally, our partners at the Middlebury Institute's Center on Terrorism, Extremism, and Counterterrorism have confirmed that fine-tuning GPT-2 on more narrow datasets tends to increase the perceived humanness of GPT-2-generated text. Fine-tuning is a key variable to take into account in the context of both human and ML-based detection.

[10] Specifically, they wrote a script to screen out generations with commonly occurring artifacts such as advertisements.

[11] Previously, the authors used best 2 out of 25 or best 3 out of 25 cherrypicking, which masked some of the differences further down the quality distribution.

[12] Note that in an earlier version of this paper, we reported findings in which the 774M model occasionally outperformed 1.5B in terms of quality. While such events occur with some probability, follow-up work has on the whole found that 1.5B is generally superior in performance than 774M.

Automated ML-based detection

Since our initial GPT-2 release, we have conducted in-house detection research on GPT-2 and seen notable work from UW, FAIR, and others.

We have seen ML-based automated detectability systems roughly fall into three categories, listed in order of complexity:

1. Simple classifiers: Uses classifiers trained from scratch to discriminate between outputs from a language model and some base "true" distribution. These can have relatively few parameters and be easily deployable.
2. Zero-shot detection: Uses a pre-trained generative model (e.g., GPT-2 or GROVER) to outputs from itself or similar models, e.g. via probabilities assigned by the model to strings of text. The model does not undergo additional training.[13]
3. Fine-tuning based detection: Fine-tunes a language model to "detect itself" with higher performance and accuracy over a range of available settings (Top-K[14], Top-P[15]).

Our Work

In May, we published a dataset of GPT-2 outputs and WebText samples [57]. In that work, we also studied discrimination between outputs and samples, where samples had an equal probability of being real or fake. And we released a simple classifier baseline that trains a logistic regression detector on TF-IDF unigram and bigram features. Using this approach, we can detect outputs from the models at Temperature = 1 at accuracies ranging from 88% at 124 million parameters to 74% at 1.5 billion parameters.[16][17] If we constrain Top-K to 40, then we can successfully detect outputs at accuracies ranging from 97% at 124 million parameters to 93% at 1.5 billion parameters. Detecting shorter outputs is more difficult than detecting longer outputs and we expect more advanced generation strategies (such as nucleus sampling[18]) could make detection more difficult than generations produced via Top-K truncation.

We also tested a simple "zero-shot" baseline using a threshold on total probability, and found that the 1.5 billion parameter GPT-2 model can detect Top-K 40 generations with between 83% and 85% accuracy. This underperforms relative to our N-gram based baseline, suggesting that it may not be easy

[13]This approach is related to the work of Gehrmann et al. on GLTR [27], which shows these probabilities to humans in a friendly interface.

[14]Top-K is a constraint that controls the number of words we consider when generating text. A Top-K of '1' would constrain GPT-2 to consistently generate its top prediction, while a Top-K of '40' means GPT-2 picks from 40 words when working out what to fill in; as we increase the Top-K we increase the variety of the generated text.

[15]Top-P controls diversity via nucleus sampling. A Top-P of 0.5 means half of all likelihood-weighted options are considered.

[16]Random accuracy in this setting is 50%.

[17]Temperature refers to controlling randomness, where lower temperatures results in less random completions. As the temperature approaches zero, the model will become deterministic and repetitive.

[18]Nucleus sampling takes samples from a variable-size set of the most probable next tokens, cut off at a certain cumulative probability, hence called Top-P.

to outperform the simplest methods. We also explore a scenario in which the adversary finetunes the model, but we are still using the original model for detection. After fine-tuning to a dataset of Amazon reviews accuracy drops to 76%, suggesting there is room for an adversary to evade detection from a static system.

Our Work: 1.5 Billion Parameter Model Detection Research

We conducted further detection research using fine-tuning, basing a sequence classifier on $\text{RoBERTa}_{\text{BASE}}$ (125 million parameters) and $\text{RoBERTa}_{\text{LARGE}}$ (356 million parameters). RoBERTa is a masked and non-generative language model that does not share the same architecture or the same tokenizer as GPT-2. Our classifier is able to detect 1.5 billion parameter GPT-2-generated text with approximately 95% accuracy. We are also releasing our detector model's code to help with detection research [58]. We acknowledge this model's dual use nature; its release intends to aid synthetic text detection research, but can allow adversaries with access to better evade detection.

The model's accuracy depends on sampling methods used when generating outputs, like temperature, Top-K, and nucleus sampling [34]. Nucleus sampling outputs proved most difficult to correctly classify, but a detector trained using nucleus sampling transfers well across other sampling methods. As seen in *Figure 1* below, we found consistently high accuracy when trained on nucleus sampling.

Figure 1: RoBERTa-Large Transferred Model Accuracy

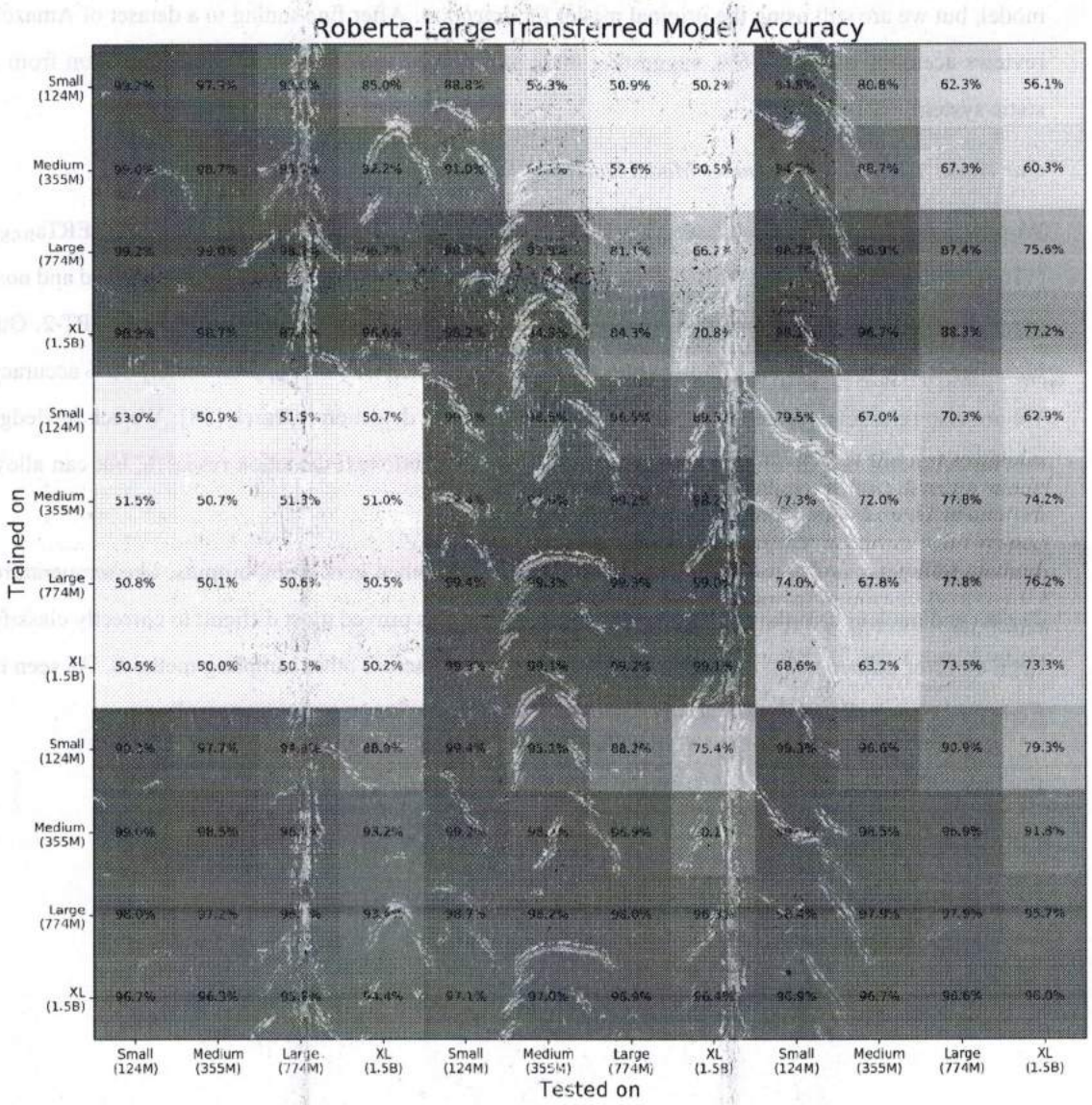

Figure 1: *The detection accuracy can be very sensitive to the sampling method of the test examples, depending on which sampling method the training examples used. To develop a robust detector model that can accurately classify generated texts regardless of the sampling method, we performed an analysis of the model's transfer performance. The 12-by-12 matrix shows the transfer accuracy with respect to the combination of four model sizes (124M, 355M, 774M, and 1.5B) and three sampling methods (Temperature = 1, Top-K = 40, and nucleus sampling with the Top-P sampled uniformly between 0.8 and 1.0). The model performs best when training samples from a larger GPT-2 model are used, which also transfers well to the test examples generated by a smaller GPT-2 model. When trained on the nucleus samples, the detector model performs well on the Temperature = 1 and Top-K 40 samples. The accuracy is obtained by testing 510-token test examples comprised of 5,000 samples from the WebText dataset and 5,000 samples generated by a GPT-2 model, which were not used during the training.*

Regardless of the detector model's capacity, training on outputs from larger GPT-2 models improves a detector's ability to classify outputs from smaller GPT-2 models well. However, the training on smaller models hinders performance when classifying larger models' outputs. Our findings imply that larger models' outputs will become more difficult to detect.

We found that fine-tuning RoBERTa achieves consistently higher accuracy than fine-tuning a GPT-2 model with an equivalent capacity. Discriminative models can be more flexible than generative models in architecture, e.g. bidirectionality, which allows them to be more powerful for detection while being less relevant to generation.[19] Our findings are in part contrary to the findings of GROVER, which suggest that the best way to defend against fake texts produced by a generative language model is the generative model itself.

We found increased accuracy in fine-tuning detection when using a mixed dataset with outputs from different sampling methods. This type of dataset helps generalize better to other sampling methods and fine-tuned outputs (e.g. Amazon reviews). We also found higher accuracy when training with random-length sequences of texts, as opposed to fixed-length texts; using random-lengths contributes to more robust classification, especially for shorter inputs. This applies most to shorter length inputs, as shorter lengths are more difficult to classify.

[19]Non-autoregressive models can also be used for generation but typically perform worse than autoregressive models.

Figure 2: Detection Accuracy With Respect to the Text Length

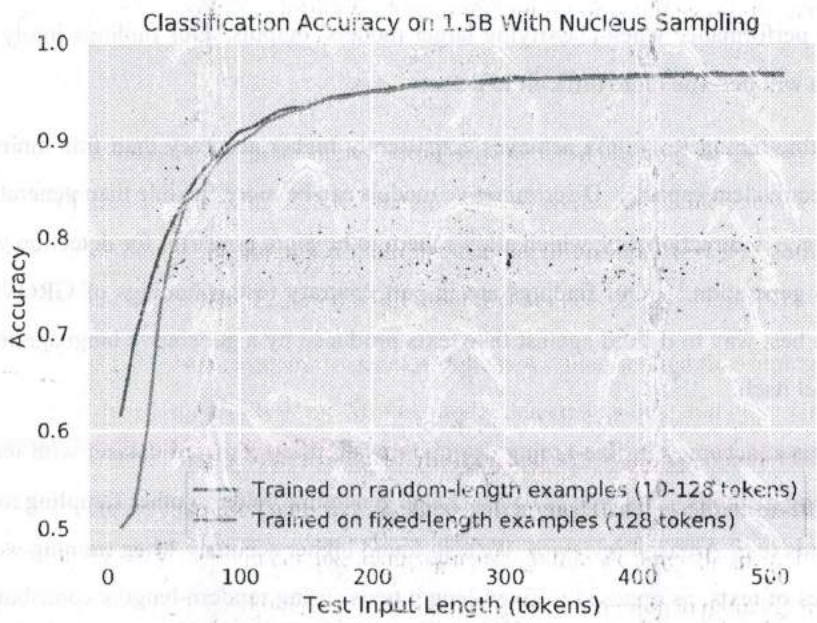

Figure 2: *The detection accuracy becomes higher for longer text, roughly surpassing 90% accuracy at 100 RoBERTa tokens (which generally translates to 70 English words). The figure also shows that training on random-length training examples has significant positive effect on the accuracy for short-length texts.*

We found smaller increases in accuracy and robustness using word dropout, where we replaced a certain percentage of training tokens with <UNK> tokens. There were similar increases in accuracy when running the detector model separately on multiple sections of an input text and gathering respective classification outputs rather than feeding the full input text at once. *Zellers et al.* [81]

Zellers et al. trained GPT-2-like systems to generate fake news, then studied fine-tuning based detection. They reported that their largest GROVER-MEGA model detected its own and other GROVER models' outputs at 92% accuracy. They also tested our 124 million and 355 million parameter GPT-2 models and found detection accuracy increased with size. Zellers et al. argued that these findings support the release of large generative models to aid in defense against misuse. While we agree there are benefits, releasing models enables misuse itself and defenses are not impenetrable. Attention to reducing tradeoffs between reducing false positives and false negatives will be needed since each has distinct implications for online platforms.

Bakhtin and Gross et al. [6]

Bakhtin and Gross et al. at Facebook AI Research study detection systems across all three classes. First, they have a baseline model somewhat similar to our simple classifier model that uses a linear "scoring function". They found this less effective than a "zero-shot" approach in their TransfBig model, a similar model to GPT-2. By using more sophisticated classifiers, culminating in one initialized from a pretrained transformer, they increased their detection rate to 93.8% in a setting with 10 negative fake examples. They also found a high degree of detection transfer from similarly sized models trained on similar data, but significant degradation when using models trained on different data.

Adelani et al. [1]

Adelani et al. found that the 124 million parameter GPT-2 could be fine-tuned to generate coherent and human-convincing fake Yelp and Amazon reviews. They tested a "zero-shot" approach based on a threshold of rare/unexpected words and used GROVER for detection [27]. Their highest detection accuracy was 97%, achieved by using GROVER on Amazon reviews.

Takeaways from the Automated Detection Landscape

While progress in automated detection is promising, existing research has yet to achieve perfect accuracy and often assumes a limited adversary. We therefore cannot draw strong conclusions about automated detection in the short run. We look forward to more work on characterizing the detection dynamics in a way that takes into account model size, training data, fine-tuning data, computational budgets for detection, sampling techniques, and other variables. Inspiration might be taken from work on the information-theoretic limits of GAN output detection [2]. In the case that such systems are insufficient, we should develop methods that involve human judgments and/or digital metadata.

Human-machine teaming

Defending against online malicious activities involves both humans and machines, using human visual interpretation skills and common sense and computers' statistical speed. Gehrmann et al. developed GLTR, a tool that automatically detects and visualizes the properties of text that correlate with the likelihood of being synthetic (e.g. out-of-context and unexpected words). Gehrmann et al. found that the use of GLTR enabled untrained humans to more accurately detect synthetic text from 54% to 72%. Notably, it is significantly easier to flag text as very-likely-synthetic, but harder to be confident that text is not synthetic. This finding supports the need for human-machine collaboration for addressing disinformation. We are also encouraged by related work in machine-manipulated images by Groh et al. [30] at MIT and the Max Planck Institute. This group found that human detection of manipulated media improves with practice. information literacy

Ippolito et al. [38] asked human raters to guess whether a passage was generated by a human or machine. They found that crowdworkers from Amazon Mechanical Turk were much worse at this task (performing at about random chance) than university students who were first walked through several examples as a group. Sampling strategy and sequence length strongly impacted detectability, with top-k samples being significantly harder to detect than those from nucleus sampling and temperature=1.0. This runs counter to the trend that we see with automatic detection systems.

Metadata-based prevention

Preventing spam, abuse, or disinformation online does not rely entirely on analyzing message content. Metadata about text, such as time taken to write a certain amount of text, number of accounts associated with a certain IP, and the social graph of participants in an online platform, can signal malicious activity. This method is used to combat attacks that use human-generated text or more simplistic and brittle forms of synthetic text generation.[20] Metadata also plays a key role in defining and justifying removing malicious content since metadata is highly complementary to the statistical analysis of text. Given this, and the difficulty of statistical detection, we expect that a wider range of platforms may need to more carefully track text-related metadata in order to be in a strong position to detect language model use (e.g. in the education system).

[20] While major tech platforms do not reveal the full details of their efforts to combat malicious activities online, there is a high level of consistency across the statements that these companies do make, in that they invariably emphasize the analysis of signals that are not a part of the sent/posted content itself. Common themes of these methods include tracking of IP addresses, tracking social graphs, and tracking the timing of messages and other events. Our conversations with experts over the past six months have broadly reinforced the impression that effective use of metadata is a key distinguishing feature of sophisticated tech platforms' efforts to combat disinformation and abuse, in combination with content-based signals as well as appropriate use of human judgment. Examples of platforms mentioning their use of metadata, include Twitter [66], Facebook [50], Google [29], and Microsoft [47]. Academic work by Yang et al. [79] also supports the view that metadata is useful in identifying social bots online, as they use features such as time zone, device information, and content deletion patterns. To be clear, we do not believe metadata is a panacea, as online malicious activity is an unsolved and perhaps intractable problem in its full generality. But the predominance today gives us some reassurance that changes to the content generation aspect of the ecosystem will not in itself be sufficient to enable major use.

4.4 Bias: Exploratory Research

Biases are reflective of both researcher choices and underlying training data. We conducted in-house tests and literature reviews in addition to external interviews and formal partnerships to study bias in language models. We are also working with the University of Oregon to develop a battery of bias probes for language models.[21] In this section we cover some preliminary of our findings from extensive literature review and bias probes.

Researchers' choices can have unintended consequences: the base language for a model biases towards outputs in that language. English-based models advantage English-speaking researchers and users relative to those from other demographics. Researchers' choice of training data can also lead to biased outputs. Training data helps define feature embeddings in the model and dataset selection conditions the model's displayed biases [51]. Biases are reinforced from a myriad of directions; occupational gender stereotypes are an example of social bias well ingrained by external influences like mass media [9]. Depending on level and field of use, language models can either reflect biases in training data or reinforce prejudices and discriminatory sentiments.

Language models like GPT-2 can be used to study how patterns in the training data can translate to biases in the outputs of large models: Societal biases expressed in the form of word connotations and context can be replicated in language models. The biases found in Internet-scale language models like GPT-2 are representative of the data on which the model was trained, which in this case was a diverse sampling of the content written in English on the Internet.[22] We have published a list of the top 1,000 sources in the 'WebText' dataset that GPT-2 was trained on to facilitate further study by researchers here [57]. We expect that internet-scale generative models will require increasingly complex and large-scale bias evaluations, the design of which will require further research and discussion.[23]

GPT-2 can generate more consistent text for a particular purpose via fine-tuning and/or "context forcing": providing GPT-2 with a long input sequence in order to more easily prime a stylistically and topically coherent output – an approach also used to trigger surprising behaviors in GROVER [24]. However, its default behavior and biases needs to be scrutinized and documented carefully by users so that they can understand and manage associated risks. We are therefore including improved documentation in our updated Github repository [59].

[21] A bias probe is an input to a model designed to elucidate the model's disposition towards producing certain kinds of outputs. We envision that a battery of such probes will be needed to comprehensively map the biases of large language models, covering issues ranging from racial and gender bias to "beliefs" in a range of conspiracy theories.

[22] For example, the top 15 domains inside the 'WebText' data on which GPT-2 was trained are (in order): Google, Archive.org, Blogspot, GitHub, the New York Times, Wordpress, the Washington Post, Wikia, the BBC, The Guardian, eBay, Pastebin, CNN, Yahoo, HuffingtonPost, Go, Reuters, IMDB, goo, and NIH.

[23] There are currently no standard methods by which to analyze bias, no established ways a model can be biased, and no unbiased researchers. Researchers and language model developers must better design frameworks and methods for bias analysis.

In Appendix C, we share some examples of both our 774 million and 1.5 billion parameter GPT-2 models' biases with respect to gender, race, religion, and language preference. We probed in these four categories due to their prevalence in our literature review and the interest in language flexibility of an English-based model, but this list is far from exhaustive and are not more or less important than other biases. In experimenting with the model, we have seen evidence that includes high associations between the word "criminal" and the male identity in GPT-2's outputs, as well as "God" with Christianity. We did not see statistically significant differences in our gender, race, or religion bias analyses between our 774 million and 1.5 billion parameter models. Language preference bias changed with the 1.5 billion parameter model, which showed more receptivity to a non-English and non-Latin script language. We shared our bias findings and gave recommendations for usage in the form of a Model Card [48] on our GitHub page [60].

Biased outputs can be useful for detecting sentiments within training data. However, as language models become more powerful and widespread, highlighting problematic biases and fine-tuning models for intended uses will be increasingly important. We encourage further bias analyses in the field of language models and encourage language model developers to test for biases in their models. There is a larger need for frameworks and standardized methods for testing for bias in language models.

5 Future Trends in Language Models

With further research, we expect language models to scale up in performance with higher output quality and accuracy. Beyond these model-level improvements, we have identified four trends to monitor in order to understand and shape social impacts of language models in a beneficial and effective manner.

Trend 1: Language models moving to devices

We can expect language models to become more widely deployed on a range of devices, given historical trends in the cost of computing power, and the current pace of efforts to move ML to perform training and/or inference on a device rather than on a server farm. For example, Hugging Face ported the 124 million parameter GPT-2 into Swift CoreML for inference on iOS devices [21].

Trend 2: More controllable text generation

Potential uses of language models will grow with developments that improve reliability and/or controllability such as new sampling methods[24], new datasets, new objective functions, and new human interfaces.

Examples of controllability include the following:

- In the GROVER model, Zellers et al. made interface modifications to introduce output controllability such that one can enter article metadata (e.g., title, author) to generate high quality outputs [81]
- The model ERNIE from Tsinghua University integrates with knowledge bases, facilitating more controllable generation than a generic language model [82].
- See et al. at Stanford and FAIR demonstrate the potential to improve chatbot performance by optimizing more directly for high-level conversational attributes such as the extent of repetition [68].
- Salesforce's CTRL model [39] improves language model controllability using what they call "control codes" to constrain model generation. Using such control codes, users can more easily steer CTRL towards generated content that is more convincing in a given context (e.g. generating content in the style of a news story [78] or a review).[25].
- Anonymous work under review at ICLR on a system called Plug and Play is also oriented in a similar direction [4].

[24] E.g. between February and now, nucleus sampling was developed by Holtzman et al. [34].

[25] Salesforce also recently published an analysis of the ethical implications of pretrained models, emphasizing the role of users and feedback processes regarding how models are used [73]

Trend 3: More risk analysis

It is currently unclear how to compare the misusability of two large language models with different performance profiles, especially when accounting for fine-tuning. Some key considerations include the time and expertise required to produce a given amount of text of a certain quality with the aid of a model versus without it, though this will change over time as technical tools evolve. GROVER generates believable news more reliably than GPT-2 due to its training data, but GPT-2's more generic training data and performance could make it easier to misuse in other ways. Beyond variations in performance at generating different styles of malicious content, different models will be more or less easy to adapt to different languages and topics. Reducing potential for misuse to zero appears difficult or impossible without sacrificing some of the flexibility that makes a language model useful in the first place. Further research and developing ethical norms are needed to take these tradeoffs seriously.[26]

Trend 4: Improved Tool Usability

Today, training and deploying of models requires knowledge of ML techniques, skill with the tools, and access to testbeds for evaluation. Steadily improved tools for interacting with language models, such as the Talk to Transformer [40] and Write with Transformer [20] interfaces, will broaden the number of actors who can use language models in a range of different ways. These improvements to tool usability will be complementary to improvements in model performance and sampling methods, and will enable an even wider array of creative applications of language models than we have seen to date.

With respect to misuse, lower-tier attackers may benefit from some of these improvements, which can reduce, but not eliminate, the gap in capabilities between lower and higher tier actors.

[26] See Whittlestone et all. [76] on the need to focus on tensions between principles in order to make progress on AI ethics.

6 Recommendations for Publication Norms in AI

There is a need for further innovation in norms, processes, and concepts for reasoning about publication-related risks in AI. We identified three recommendations for AI practitioners to build capacity in navigating responsible publication in AI.

Recommendation 1: Build frameworks for navigating tradeoffs

While the staged release method seeks to reduce harms and maximize benefits, we found weighing both pre-publication was difficult and there is an urgent need to develop principled decision-making frameworks.

In creating frameworks, systems that have an impact outside the AI community should undergo interdisciplinary analyses among researchers and broader society.

In March, OpenAI and the Partnership on AI, alongside other members of the AI community, co-hosted a discussion on publication norms. In June, OpenAI began work with the Partnership on AI on a project relating to publication norms in AI research; while this project is as-yet unpublished, it gathers the views from companies, organizations, and people differently affected by artificial intelligence to present key considerations and ideas for developing responsible publication norms as a community

Recommendation 2: Build infrastructure for distributed risk analysis

yes, always have risk analysis & mitigation attached.

We aimed to prevent premature publication while enabling other researchers to contribute to risk analysis. Working with prospective partners, we designed legal agreements that balanced both parties' interests, minimizing red tape and logistical burdens. We saw Zellers et al. take a conceptually similar approach with GROVER, giving early access to researchers. We have had productive discussions with them and others about improving processes for distributed risk analysis. Our legal negotiation process and subsequent learnings about GPT-2 demonstrate that there is no standardizable model sharing approach. We provide a template agreement in Appendix A to help organizations develop appropriate processes in this area.

We identify areas to improve in legal and technical infrastructure for model sharing below [62]:

- **Scalability:** Currently, agreements require fine-detail discussion and negotiation. An alternative approach might be a system in which participants are vetted once and can subsequently access more than one model under the same terms.
  - Related approaches are used in other contexts such as genomics data sharing [53].
  - Zellers et al. [80] also note the challenge of scalability and discuss other possible approaches.
- **Security:** There is a tradeoff between the number of partners and the likelihood of a model being prematurely released, accounting for hacks and leaks.
- **Fairness:** The high cost of compute used in powerful models like GPT-2 raises concerns about accessibility and equity in future AI research [13]. Private model sharing should not excessively harm researchers with limited computing resources, and conflicts of interest related to model sharing should be avoided in commercial contexts.

Recommendation 3: Build communication channels across organizations

as they did — but need to consider security leaks, national interests

Research results are often kept private until the associated paper is published. Private results hinder coordination, especially for release; for example, we were largely unable to retrieve statuses of replication efforts. The norm of privacy around unpublished research holds legitimacy, as seen in non-disclosure agreements, but robust communication channels between AI organizations will be needed in the future. For example, prior to first announcing GPT-2, we were unsure whether and how quickly other labs would eventually develop and publish similar systems. Since the impact of an individual publication decision often depends on others' publication decisions, we encourage AI labs to experiment with their approaches to interorganizational communication.

Conclusion

We saw evidence of positive applications and minimal evidence of planned misuse, and research into detection properties and biases, in addition to collaborations among researchers and cautious approaches to publications. These findings as part of our staged release and partnerships processes gave us confidence to release our1.5 billion parameter GPT-2.

We saw researchers and engineers apply GPT-2 for a range of positive uses, giving us reason to expect similarly beneficial uses with larger models. Furthermore, our analysis of the landscape of malicious actors has led us to believe that our staged release process will primarily affect the low and middle ends of the actor distribution, with little evidence of large-scale misuse. However, we also expect that the skills and resources required for using language models, both beneficially and maliciously, will decrease over time. We therefore recommend the AI community build frameworks for navigating tradeoffs, infrastructure for distributed risk analysis, and communication channels across organizations.

Beyond language, researchers at OpenAI and elsewhere are training increasingly powerful generative models on a range of media, including images, video, and audio. While we expect lessons from GPT-2 to inform some decision-making in other large-scale generative models (e.g. the concepts of staged release and partnership-based model sharing), there will be more novel challenges and opportunities. We hope GPT-2 as a case will help the AI community navigate publications in omni-use AI research.

Acknowledgements

We thank the following individuals for feedback on earlier versions of this document:

Gillian Hadfield, Haydn Belfield, Cullen O'Keefe, Clément Delangue, Sarah Kreps, Miles McCain, Rowan Zellers, Emily Alsentzer, Nathan Benaich, Jason Blazakis, Sam Bowman, Sebastian Gehrmann, Chip Huyen, Daphne Ippolito, Carson Kahn, Subbarao Kambhampati, Daniel Lowd, Andrew Mauboussin, Stephen Merity, Luke Muehlhauser, Robert Munro, Alex Newhouse, Larissa Schiavo, Adam Shostack, Lavanya Shukla, Ravi Srinivasan, Charlotte Stix, Michael Littman, Cody Wild, Rebecca Crootof, Vanya Cohen, Aaron Gokaslan, Connor Leahy, Mona Wang, Jeremy Gillula, Myle Ott, and Lav Varshney.

Any remaining errors or omissions are the authors' responsibility alone.