A COST ANALYSIS OF GENERATIVE LANGUAGE MODELS AND INFLUENCE OPERATIONS

Micah Musser
Center for Security and Emerging Technology
Georgetown University
Washington, D.C.
Micah.Musser@georgetown.edu

ABSTRACT

Despite speculation that recent large language models (LLMs) are likely to be used maliciously to improve the quality or scale of influence operations, uncertainty persists regarding the economic value that LLMs offer propagandists. This research constructs a model of costs facing propagandists for content generation at scale and analyzes (1) the potential savings that LLMs could offer propagandists, (2) the potential deterrent effect of monitoring controls on API-gated LLMs, and (3) the optimal strategy for propagandists choosing between multiple private and/or open source LLMs when conducting influence operations. Primary results suggest that LLMs need only produce usable outputs with relatively low reliability ($\sim 23\%$) to offer cost savings to propagandists, that the potential reduction in content generation costs can be quite high (up to 70% for a highly reliable model), and that monitoring capabilities have sharply limited cost imposition effects when alternative open source models are available. In addition, these results suggest that nation-states—even those conducting many large-scale influence operations per year—are unlikely to benefit economically from training custom LLMs specifically for use in influence operations.

***K*eywords** Influence Operations · Language Models · Cost Modeling

1 Introduction

For the past several years, experts have speculated that newly emerging large language models (LLMs) may be used by malicious actors to generate divisive, misleading, or false information for the purposes of social manipulation. [3, 4, 8, 18, 29, 36, 47, 52] Organizations releasing such large language models have explicitly acknowledged this as a misuse risk, [48, 61] and some major players advocate for "best practices" on limiting access to large language models that include "publish[ing] usage guidelines" and "build[ing] systems and infrastructure to enforce usage guidelines." [11] However, other organizations and commentators have expressed skepticism that influence operations would benefit from using language models to produce content, potentially because they still require human curation or because the costs of generating disinformation content are already extremely low. [6, 30, 31, 40, 54] Such uncertainty has resulted in calls to explicitly evaluate the costs of conventional influence operations as compared to AI-enabled ones, as illustrated in the following quote from a 2020 workshop:

> [M]odels like GPT-3 can be used to create false, misleading, or propagandistic essays, tweets, and news stories *de novo* . . . [W]hile automated generation of disinformation may be feasible in principle, human labor may still be more cost-effective for such purposes. Others disagreed, and saw automated generation as much more cost-effective than training and paying humans to generate disinformation. Participants agreed that empirically investigating the economics of automated vs human generated disinformation is important. [57]

Despite this interest in the economics of influence operations, the topic remains underexplored, no doubt due in large part to the difficulty of assessing the economics of presently existing and highly secretive influence operations. To this author's knowledge, only one public attempt has been made to actually model the costs and tradeoffs facing influence operators deciding whether or not to make use of AI systems. [22] This research primarily addresses the economics of deepfaked visual and audio content, with a focus on whether or not a technological "arms race" between detection systems and malicious actors is likely to happen.[1]

This paper attempts to explore two related but different questions. First, what are the potential economic benefits of using language models to produce disinformation content, relative to human authorship? And second, what is the economic value of one possible policy intervention designed to reduce the risk of automated disinformation generation, namely, the use of monitoring controls on privately-held models? To make progress on these questions, this paper attempts to models the costs of *content generation* for an influence operator in various situations, including the use of a public (open source) language model, a private, monitored language model, or a manual campaign. I emphasize that this analysis focuses very specifically on the costs of content generation, which is only one part of the disinformation pipeline and may be—for some operations—less costly than other requirements facing operators, such as maintaining an infrastructure of inauthentic accounts or identifying the appropriate channels for distributing content. [17, 51]

The paper is organized as follows: Section 2 discusses a simple base scenario: how much could the use of a language model save if the lanugage model required no human curation and could be deployed fully autonomously? Section 3 then analyzes the more likely scenario of human-machine teams, where human curators review and approve model outputs instead of writing content themselves. In section 4, I then consider the cost imposition that could be generated by the use of monitoring controls on a single AI model available to would-be influence operators. Section 5 considers the value of monitoring controls under circumstances which permit operators to choose between the use of multiple models, including fully public ones. While all analyses through Section 5 focus only on marginal costs, Section 6 expands this to include an analysis of the fixed costs associated with different methods for accessing a language model. Section 7 further analyzes the robustness of the results from the preceding sections, and 8 concludes with a discussion of the implications of this research.

This analysis is strictly focused on the use of language models to generate short social media posts (which I will often refer to as "tweets" for the sake of focusing attention on a specific use case with relatively constant output lengths, though there is nothing platform-specific about this analysis). The model can generalize to other types of language content as well, such as news articles or blog posts. Newer text-to-image models have sparked analogous concerns about disinformation uses [5, 59]; this model in principle can also generalize to non-text-based forms of content generation, though the interpretation of some key parameters must change.[2] I emphasize that the model and its usage here are meant to be "first attempts" at explicitly modeling the cost decision facing a malicious actor who is deciding whether to make use of a large language model; it has several key limitations (see Section 8), but nonetheless may hopefully serve to inspire further refinements.

2 Fully Automated Content Generation

For an "ordinary" campaign, where all content is manually authored by humans, let L represent the labor productivity of human authors (measured as outputs per hour) and w represent the hourly wage of human authors. For simplicity, I treat both of these variables as constant over the full course of an arbitrarily long influence operation, which implies that the marginal cost of an additional output is constant and equal to w/L.[3] The only real difficulty when framing a manual campaign in these terms is to esimate these two values.

[1] For some discussion of the "arms race" dynamic in synthetic content, see [34]. Other research, such as [32], has attempted to model the impact of AI-enabled phishing campaigns, but without examining whether the use of such AI tools is therefore cost effective for malicious actors.

[2] For instance, the cost to review an output might, when considering a text-to-image model, include touch-up work done by a human designer to finalize model outputs for posting.

[3] One objection to this framing is that in real influence operations, the cost of a marginal output declines over time due to the widespread use of time-saving tactics such as "copypasta," as reported for instance in [12]. There are two reasons to think that this objection need not require a non-linear estimation of manual costs. First, w/L can be thought of as an **amortized** cost per output, and not as an **extrapolated** cost per output based on assumptions about how long it would take to write posts *de novo*. The method for estimating these parameters, which relies on historical reporting about monthly wages as well as expected output per shift for operators, aligns more readily with this interpretation of w/L as an amortized rate and not an extrapolated one. Second, while human propagandists do frequently resort to copypasta and other tactics, language models are less liable to do so. Insofar as copypasta remains a key way of identifying coordinated inauthentic activity, the use of a language model would therefore represent a quality improvement over human authorship not captured by straight comparisons of the cost per output of either a human or a language model. I am inclined to think that, due to reason one, the model presented here accurately captures the **cost** differential between a manual operation and an AI-augmented one, but also that—due to reason two—it is liable to understate the **quality** improvement that arises from switching to the use of a language model once doing so becomes cost-effective.

Estimates of either the wages paid to disinformation authors or the productivity of such works are hard to find, though some scattered pieces of information do exist, primarily in the context of Russian influence operations. (Because wages and labor productivity likely vary widely across campaigns conducted in different regions of the world, the specific estimates produced by this model can be thought of as reflective the value of LLMs specifically for Russian influence operations, though see footnote 7.) In 2018, *BuzzFeed News* reported that the Internet Research Agency (IRA) had posted job ads in 2014 and 2015 for "social media specialists" and "content managers" paying roughly $2.86–9.53 per hour.[4] [33] Reporting from the indepenent St. Petersburg-based publication *Fontanka* in 2022 surfaced more information: a reporter who successfully interviewed for a job with "Cyber Front Z" (which appears to be linked in some way to the IRA [16]) was offered a job that would pay $1.41–2.78 per hour.[5] [27] In addition, another (older) article from 2014 in *BuzzFeed News* implies an average hourly wage for IRA employees in the range $3.62–5.44.[6] [50] Though this source does not contain direct information about salaries, the fact that it falls within the overall range suggested by the other two sources is encouraging.

Some of these sources also contain information about the expected output of content generators working for the IRA. The *Fontanka* report noted that employees at Cyber Front Z were expected to write 200 comments on social media posts per shift, or somewhere in the range of 20–25 comments per hour, depending on the length of a shift. But the 2014 *BuzzFeed* article about older IRA campaigns suggests thats operators managing Twitter accounts were only expected to tweet 5–6.25 times per hour.

In the following models, I use Monte Carlo simulation to estimate both w and L, treating both as random uniform distributions over the full range given by the above estimates. The smallest possible cost of a marginal output w/L given these parameter ranges is therefore $0.06, the largest possible cost is $1.91, and the expectation of the marginal cost is $0.44.[7]

In the alternate case where an influence operation employs a language model to generate content fully autonomously, the only marginal costs associated with content generation are those required to run inference on a model. For its largest, most powerful language model, OpenAI curently charges $0.00006 per token, while Cohere charges an even lower $0.000015 per token for generation tasks. [10, 45] Since I am generally considering tweets or comments on social media as the standard type of content in this threat model, I estimate the average token length of outputs at around 40 tokens, in which case the marginal cost for an additional output from a major company model is likely in the range $0.0006–0.0024. If, alternatively, a threat actor uses a public model which requires them to set up and maintain their own compute infrastructure, these costs may be higher, but it seems reasonable to estimate that, for any reasonably large operation, an operator could keep inference costs within an order of magnitude of the costs offered by major companies.[8] The estimated values for the marginal inference cost IC of an additional AI output therefore fall roughly in the range of $0.0006–0.024.

Given these estimates, a threat model of **pure** automation will always have lower marginal costs than that of a manual influence operation. This is not surprising. In addition, if an operator must expend fixed costs FC to acquire a working model (whether that means training it from scratch, stealing it, fine-tuning it for an operation, or even just familiarizing

[4] The specific figure from [33] is for 40,000 rubles per month for two different jobs posted sometime in either 2014 or 2015. The lower bound of this estimate comes from converting rubles to dollars at the lowest conversion rate within those years, converting to 2022 USD, and then assuming 240 hours of work per month (10 hours of work, 6 days a week, for 4 weeks). The upper bound comes from converting rubles to dollars at the highest conversion rate within those years, converting to 2022 USD, and then assuming 160 hours of work per month (8 hours of work, 5 days a week, for 4 weeks). The reason for the wide spread is primarily that the value of the ruble fell dramatically in late 2014.

[5] The level of variation is again due to the use of 160 hours of work per month as a high-end estimate and 240 hours of work per month as a low-end estimate, as well as the fact that the value of the ruble fluctutated significantly in the 10 days between Cyber Front Z's hiring call and the publication of *Fontanka*'s report.

[6] The *BuzzFeed* article places the total estimated budget of the IRA in 2014 at $10 million, with half "earmarked to be paid in cash" (likely for employee salaries, of which the organization had 600 at the time). If we assume that these employees worked between 160 and 240 hours per month, and that all cash-earmarked funds were paid out as salaries, then the average hourly wage for IRA employees in 2014 would have fallen in the range $3.62-5.44 after adjusting for inflation. This may slightly overstate the figure for employees who were tasked with content generation, who may have earned lower wages overall than other types of employees.

[7] Conveniently, although these numbers were taken from a variety of sources related to Russian propaganda efforts, an expectation of $0.44/post happens to align nicely with the notion of the "50 cent army," the traditional term used for Chinese propagandists who were assumed to be paid roughly $0.50 for each post they wrote. Although [25] questions this estimate and suggests that most Chinese propagandists are salaried bureaucrats, the authors do not provide an alternate way of estimating the effective cost to the government of each post produced by these bureaucrats.

[8] While major companies like OpenAI and Cohere benefit from very large economies of scale, they also deploy much larger models than a propagandist would be likely to run on local equipment, see Section 6. In addition, final estimates do not depend heavily on values of IC (see Section7), allowing for some looseness in estimation here.

one's staff with the model's capabilities), then the model pays for itself after $\frac{FC}{w/L-IC}$ outputs. With expected values $E(\frac{w}{L}) \approx 0.44$ and $E(IC) \approx 0.01$, then the use of an AI model would pay for itself after a campaign of size $2.33FC$.

3 Human-Machine Teams with Unrestricted AI Access

With current models, it is unlikely that in **most** cases, an operator would choose to run a purely automated campaign.[9] For most campaigns, especially those where the consistency and quality of posts matters heavily to the campaign's overall success, a human-machine team is a more realistic scenario. For the purposes of this paper, I imagine that a human-machine team operates in the following way: a language model is tasked with outputting content which is subsequently reviewed by a human prior to posting. The human must approve an output (perhaps with some light editing) before it is posted online.

To incorporate an operation along these lines into this model, I introduce two additional parameters. First, let α represent some constant proportion indicating the efficiency gain associated with using a language model, such that αL represents the total number of posts a human can generate and review in an hour. And second, let p represent the proportion of outputs from a language model that are usable for an operator's campaign (or that will be usable after a light edit during the review process).[10] Then the cost of producing a marginal output using a human-machine team can as be modeled as a constant, with this strategy being cheaper than paying a human to write a marginal output whenever the inequality

$$\left(\frac{w}{\alpha L} + IC\right)\frac{1}{p} < \frac{w}{L} \tag{1}$$

obtains. Note that, because the inference costs of running a model are generally dwarfed by labor costs, this inequality loosely approximates to the inequality $\alpha > 1/p$, which states that whenever the speedup in a human's ability to produce and review AI generations (compared to writing them manually) is greater than the number of AI generations necessary to find a "usable" output, we expect the marginal cost of an output from a human-machine team to be cheaper than the marginal cost of a human writing an additional output.

Choosing an appropriate value for α is one of the more difficult tasks associated with parameter estimation in this model. Although some economic studies on the labor impacts of large language models have begun to emerge, [7, 9, 14, 24, 28, 56] they are mostly of limited usefulness. [9] and [28] speculate that large language models will enable efficency gains for human workers but do not measure such gains, while [14] analyzes worker exposure to large language models but not efficiency impacts of the models. [7] estimates a 14% efficiency improvement among call center workers using large language models, which corresponds to an absolute minimum estimate of $\alpha \approx 1.14$.[11] However, this value reprsents an an aggregate efficiency gain across all worker tasks, not a relative speedup specific to a discrete task, which is necessary to estimate the savings that large language models could offer propagandists for the specific task of content generation. [24] analyzes efficiency gains for a specific code completion task and provides an absolute minimum estimate of $\alpha \approx 2.27$, while [56] estimates α at 4.26.[12]

[9] However, note that if the goal of a campaign is distraction, such that the quality of individual posts does not matter to the operator, pure automation may be a perfectly workable strategy for existing language models. See [25] for an analysis along these lines in the context of Chinese influence operations.

[10] Note that α and p will be inversely related if the actual underlying capability of a given model remains constant: reviewers can choose to spend more time editing potential AI outputs or engaging in careful prompt engineering, thereby increasing the proportion of outputs that are considered "usable" at the cost of reducing α, or they can simply make binary yes-no rulings on potential outputs, which increases α at the cost of reducing p. For any given combination of model and campaign, there is likely some optimal level of investment that reviewers should make in each output, but this would be hard to predict *a priori*. This model attempts to handle this ambiguity by sampling from a relatively wide range of values for α while treating p in most places as an entirely free-floating variable. But it is important to emphasize that readers trying to imagine plausible values of p for existing models should **not** interpret this parameter as corresponding only to the proportion of outputs from language models that are perfectly suited for use in an influence operation with no editing or prompt engineering whatsoever.

[11] In my model, the efficiency speedup associated with the use of an LLM is disaggregated into an increase in the rate at which humans can generate and review outputs, compared to manually writing them (α), offset by the percentage of outputs that are actually usable (p). This disaggregation is necessary when evaluating operator costs, because inefficiencies caused by lowering p generate increased inference costs, while inefficiences caused by lowering α do not. In addition, the disaggregation separates efficiency gains into a parameter specific to a given human-task pair (which I estimate using Monte Carlo methods), and a parameter specific to a given model-task pair (which I primarily treat as a free-floating variable). Because $L_{\text{AI}} = \alpha p \cdot L_{\text{Human}}$, an overall productivity gain of 14% corresponds to a value of $\alpha = 1.14$ when p is equal to its maximum value of 1.0.

[12] [24] measures only the observed efficiency improvement on a given coding task and does not estimate p, similarly to [7]. [56], however, estimates a total efficiency gain of 6% from the use of AI **and** finds that 25% of AI suggestions were accepted by coders, implying a value of $\alpha = 4.26$. [39] does not estimate the total efficiency gain of using AI to assist with coding tasks but does observe

For the purposes of this paper, I randomly sample values for α uniformly from the range $[2, 10]$, indicating a wide range of uncertainty about how much faster it would be for an operator to generate and review outputs instead of manually writing them.[13] Figure 1 simulates 10,000 possible parameter values for each parameter except p and plots the cost savings of a marginal output that could be gained from switching from manual authorship to a human-machine team as a function of p. Whenever p is above 0.23, we expect that, on average, there will be some positive cost savings associated with using a human-machine team instead of a fully manual operation. If half of the outputs of the AI model are usable, we expect that on average a human-machine team will cost \$0.24 less per output than a manual operation; since the per-output cost of the manual operation has an expectation of \$0.44, this represents a 55% reduction in costs.

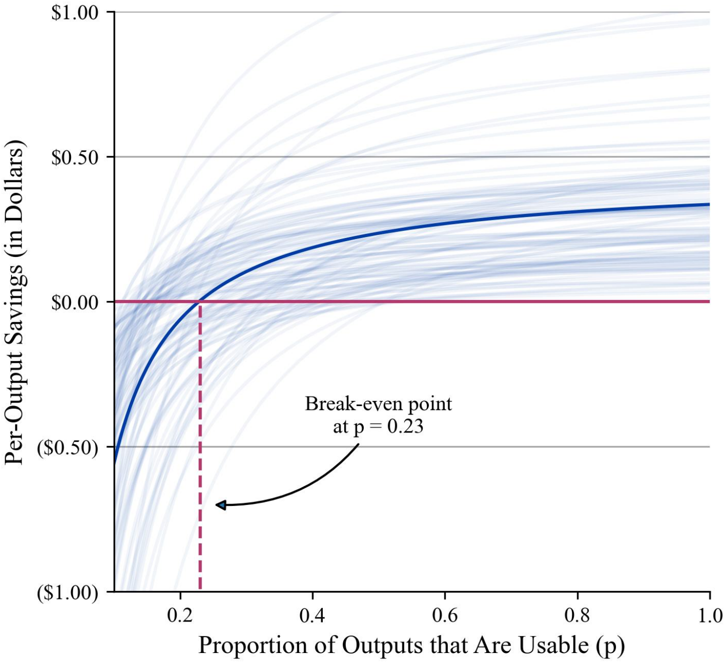

Figure 1: Simulated Per-Output Cost Savings as a Function of p

Over a sufficiently long campaign, a per-output savings of 55% can add up to relatively significant amounts. Figure 2 shows cumulative savings as a function of both p and campaign length, up to 10 million tweets. (Solid lines designate savings of one million dollar increments, with dashed lines designating increments of \$500,000; savings are calculated as the mean savings across 10,000 Monte Carlo simulations.) It is worth emphasizing that for multiple nation-states, the posting of several million tweets to Twitter is an entirely realistic goal in the medium-term. Based solely on publicly released datasets of coordinated inauthentic activity on Twitter, actors affiliated with the following countries all appear to have posted multiple millions of inauthentic tweets prior to December 2021: Serbia (17M), Saudi Arabia (17M), Turkey (15M), Egypt (7M), Iran (5.5M), Russia (5M), the United Arab Emirates (4.9M), China (3.9M), Venezuela (3.8M), and Cuba (2M).[14] These estimates are based only on infrequently released Twitter data partially covering October 2018–December 2021 (with long gaps between some releases), and is therefore likely a major undercount not only of inauthentic state-affiliated activity on Twitter specifically, but even more so of state-affiliated influence operations generally.

that roughly 22% of AI-generated suggestions were accepted by coders in a large-scale deployment. Because this is consistent with the acceptance rate observed by [56], it is likely more accurate to say that [24] implies a value of α between 9 and 10.5 when assuming a corresponding value for p of roughly 0.22–0.25.

[13] The ends of this range do, however, correspond roughly to the minimum and maximum estimates of α provided or implied by productivity improvement observed in specific code completion tasks with the use of AI models as discussed above.

[14] See [60]. These figures were calculated based on the file sizes of "Tweet Information" files for each campaign, which contain metadata about tweets. As a baseline, the 353MB file corresponding to the Russian campaign released in June of 2020 contains 1.04 million tweets, suggesting that roughly 340MB of data corresponds to 1 million tweets. Note that one 4.2GB file was attributed to a joint campaign between Saudi Arabia, the United Arab Emirates, and Egypt; for simplicity, I simply divided the (imputed) number of tweets in this campaign evenly across all three countries, though it is likely given the objectives of the campaign that a disproportionately larger number of tweets came from Saudi Arabia.

In short, if influence operators had unrestricted access to language models capable of producing usable text at least 50% of the time, this model predicts that an operator could save upwards of $2 million over the course of a 10-million tweet campaign, with an estimated reduction in per-output content generation costs of 55%. Moreover, based on public information about Twitter takedowns, there is a meaningful number of nation-state actors who appear likely to produce >10 million tweets (or the equivalent amount of text on other platforms) in the near- to medium-term.

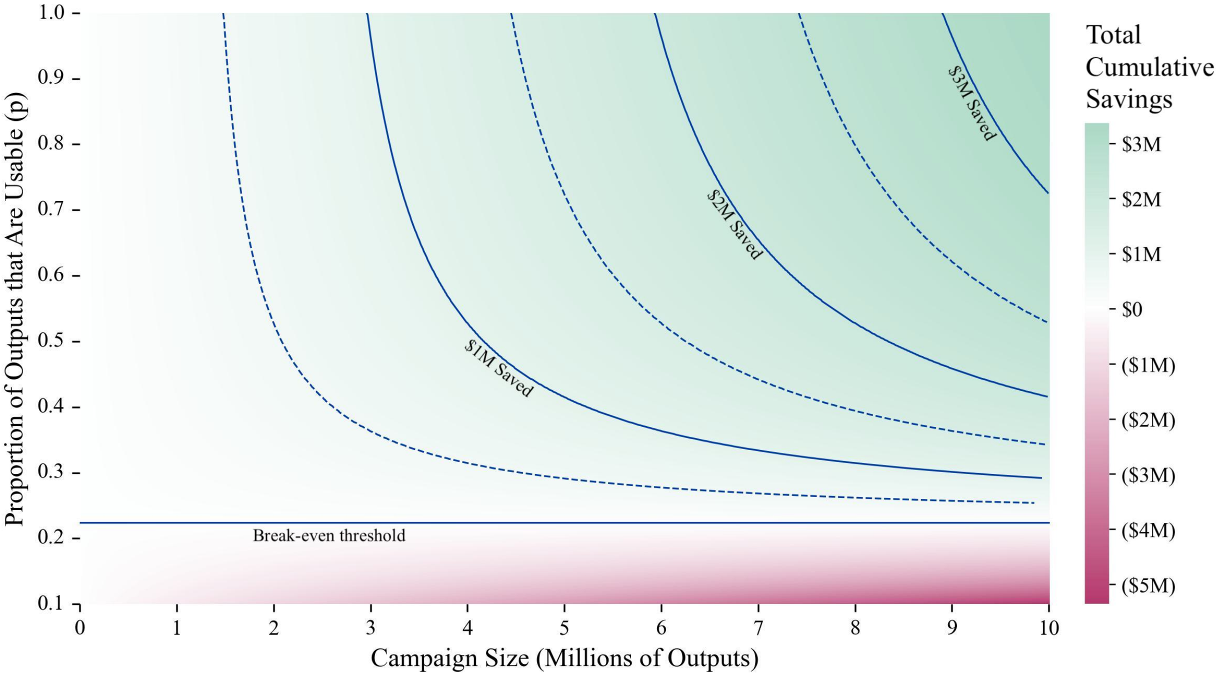

Figure 2: Cumulative Savings as a Function of Campaign Length and p

4 Monitoring Controls on AI Models

Not all language models can be accessed by potential propagandists without restrictions. In the case of ChatGPT, for instance, the model itself continues to be held privately by OpenAI, with users of ChatGPT being required to make an account in order to access the model.[15] Early beta users of GPT-3 were required to provide a description of their intended uses of the model prior to being granted access, but roughly eighteen months after the model's announcement, OpenAI removed the waitlist and allowed more immediate access to the model; OpenAI followed a similar trajectory with its text-to-image model DALL·E 2 after only five months.[43, 44] While the model has become increasingly available to anyone to use, the fact that it remains behind a closed API makes it possible for OpenAI to monitor user interactions with ChatGPT. Users who are deemed to be deliberately generating harmful content could have their access to the model revoked via a revocation of API access tokens, IP address blocking, or some other measure.

Can such monitoring controls impose meaningful costs on propagandists attempting to use language models to conduct large-scale influence operations? While blocking a user account or IP address imposes a penalty on a malicious actor, propagandists can generally create a new account or use a new IP address to continue accessing the same model, at which point the detection process must restart. In other words, if there is a roughly constant rate of detection per output λ, and each detection D incurs a penalty P and resets the clock for the next detection, then the costs imposed by monitoring controls over a given campaign length can be modeled as a random draw from a Poisson distribution of detections, multiplied by the penalty for each detection. Then the costs C of a campaign of size n will be equal to the

[15] Note that there are many finetuned, downstream applications of ChatGPT that may not require end-users to sign up for API access. However, the creator of the downstream application must themselves maintain API access to ChatGPT, and could potentially have such access revoked if their users appear to be abusing their indirect access to the original model.

minimum of either the manual cost of producing content, or the cost of using a language model to generate content plus the costs of evading detection:[16]

$$C(n) = \min\left(\frac{nw}{L}, \quad \frac{n}{p}\left(\frac{w}{\alpha L} + IC\right) + P * D \sim \text{Pois}\left(\lambda \frac{n}{p}\right)\right) \tag{2}$$

The penalty paid for a detection could conceivably be quite low, if a human must simply generate a new email account and sign up for the API again. Even so, doing so may generate friction costs as the human switches between reviewing outputs to creating a new account. Companies may also adopt relatively more stringent deterrance methods, for instance by requiring proof-of-humanness to sign up for an API. I—perhaps generously—imagine that each detection could require between 0.5 and 2 hours of a worker's time to evade before an operation can resume. This means that $P \sim U(0.5w, 2w)$, where w itself is sampled from a uniform random distribution.[17]

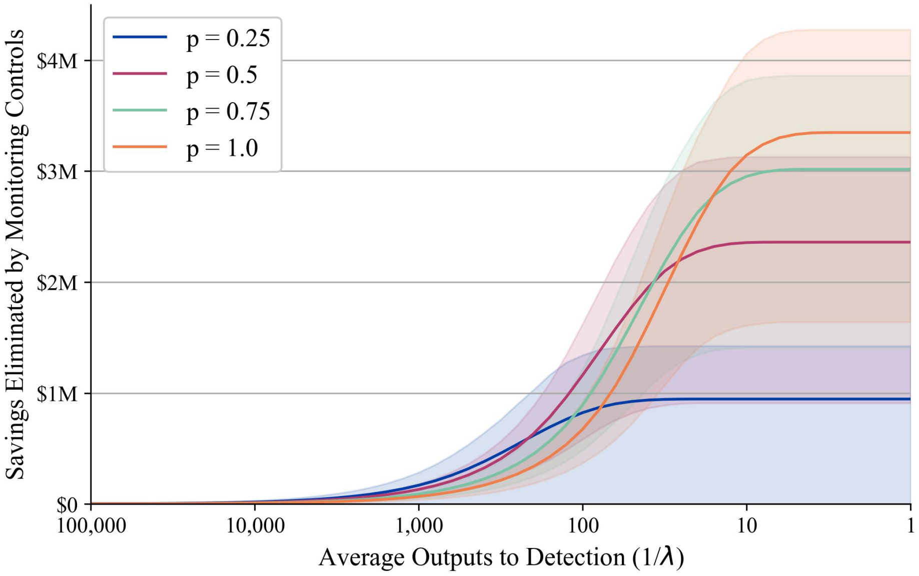

Figure 3: Penalties Imposed by as a Function of Monitoring Efficacy, for Varying Levels of p

Figure 3 shows, for four possible values of p, how improvements in detection capabilities alter the costs imposed by monitoring controls. The figure suggests that improving detection capabilities has different effects over three general phases:

1. If the malicious user can, on average, produce roughly 1,000 or more outputs without detection, the monitoring controls impose minimal costs. Improvements in the ability to monitor the model do not substantially alter the cost calculus that propagandists perform.
2. As capabilities improve from one detection per 1,000 outputs to about one detection per 100 outputs, costs imposed on propagandists increase by roughly similar dollar values regardless of the underlying model's capabilities.
3. Somewhere between one detection per every 100 outputs and one detection per every ten outputs, the monitoring controls impose costs equivalent to the difference between a manual campaign and an AI-augmented one. At or above this threshold detection capability, the propagandist prefers to use a manual campaign, and further improvements in detection capabilities impose no additional costs. The total costs imposed by monitoring controls are significantly greater for better-performing models (as the potential savings from the use of the AI model were originally much larger), but better detection capabilities are required to fully impose such costs.

[16] Note that n here refers to the number of *usable* outputs that have been produced by the model, but since all outputs (usable or not) contribute to the model owner's ability to detect malicious use, n must be divided by p in equation 2 to account for unusable outputs that nonetheless contribute to the eventual detection of the malicous use.

[17] In dollar terms, $E[P] = \$6.84$.

The shaded regions in Figure 3 represent the interquartile range of possible outcomes across 10,000 simulations. There is clearly an enormous deal of variation in the dollar value of costs imposed by monitoring controls, which is further analyzed in Section 7. The general transition phases, however, are consistent across nearly all simulations: a detection capability that revokes access once per every ten outputs completely eliminates the incentive to use a model, while detections once every 100 to 1,000 outputs still impose meaningful costs.

5 The Value of Monitoring when Public Models Are Accessible

The preceding section imagines that a propagandist must decide whether to produce content using a private, monitored language model or a manual process of human authorship. This might be plausible if there were a single (API-restricted) language model available to propagandists; in the real world, however, many language models have proliferated rapidly. [18, 53] In this section, I instead imagine that a propagandist can choose between the use of two different language models, where model 1 is an API-accessible model with monitoring controls in place, and model 2 is a public model instead. For now, I examine only the **variable costs** associated with either generation strategy, though in the next section I briefly discuss the fixed costs associated with downloading, finetuning, and running a public model. Assuming that both the public and the private model satisfy inequality 1, the propagandist would prefer to use the private model 1 so long as the condition

$$\left(\frac{w}{\alpha L} + IC_1\right)\frac{n}{p_1} + P * D \sim \text{Pois}\left(\lambda\frac{n}{p_1}\right) < \left(\frac{w}{\alpha L} + IC_2\right)\frac{n}{p_2} \tag{3}$$

is satisfied. To make this equation slightly more manageable, we can assume that the inference costs are the same regardless of model.[18] From the propagandist's perspective, where D is unknown at the start of the campaign, we may also substitute $D \sim \text{Pois}\left(\lambda\frac{n}{p_1}\right)$ with $E[D]$, which is just $\lambda\frac{n}{p_1}$. Then, with some rearranging, inequality 3 becomes:

$$P\lambda < \left(\frac{w}{\alpha L} + IC\right)\left(\frac{p_1 - p_2}{p_2}\right) \tag{4}$$

Inequality 4 states that the propagandist's expected marginal costs from relying on a private, monitored language model are lower than those of the public model only if the penalty per detection times the detection rate is lower than the marginal cost of reviewing an output, times the percentage performance improvement that the private model offers relative to the public one.[19]

Let $\hat{p}$ represent the threshold performance of an AI model at which it becomes cost-effective to use the model, relative to a manual campaign. Then there are four relevant scenarios that determine the propagandist's cost-optimal strategy:

1. If $p_1 \leq \hat{p} \wedge p_2 \leq \hat{p}$, the propagandist prefers to use a manual campaign regardless of any monitoring controls on the private model;
2. If $p_2 > \hat{p} \wedge p_2 > p_1$, the propagandist prefers to use the better-performing public model regardless of any monitoring controls on the private model;
3. If $p_1 > \hat{p} \wedge p_2 \leq \hat{p}$, the propagandist prefers to use the private model, but will fall back to the use of a manual campaign if monitoring controls impose sufficient costs; and
4. If $p_2 > \hat{p} \wedge p_1 > p_2$, the propagandist prefers to use the private model, but will fall back to the use of the public model if monitoring controls impose sufficient costs.

For each pair (p_1, p_2) that satisfies either condition 3 or 4 above, it is possible to estimate the minimum detection capability $\hat{\lambda}$ that imposes sufficient costs to deter the propagandist from using the API-gated model. For condition 3, this value can be estimated using equation 2, and for condition 4, it can be estiamted using equation 4.[20] Since further improvements in detection impose no additional costs after a propagandist has resorted to their fallback strategy, the

[18] This decision is justified by the fact that inference costs are generally dwarfed by labor costs in this model; see Section 7.

[19] Note that if the public model is actually better-performing than the private model, the right-hand side of inequality 4 will be negative. Since the left-hand side is necessarily positive, this means that the private model is never preferred to the public model when considering only **variable** costs. However, as Section 6 briefly discusses, it is possible for a propagandist to prefer a worse-performing private model over a public one if the fixed costs associated with the public model are sufficiently high.

[20] See Appendix A for equations used to calculate $\hat{\lambda}$, including when the propagandist must pay fixed costs to access and/or finetune a public model.

maximum cost imposition of monitoring controls over a campaign of length n can further be estimated as $P\hat{\lambda}\frac{n}{p_1}$. Figure 4 shows, for all values of (p_1, p_2), the optimal strategy pursued by the propagandist, the detection capability needed to make the propagandist indifferent between the use of the private model and the relevant fallback option, and the costs imposed by such a detection capability over a ten-million-tweet campaign.

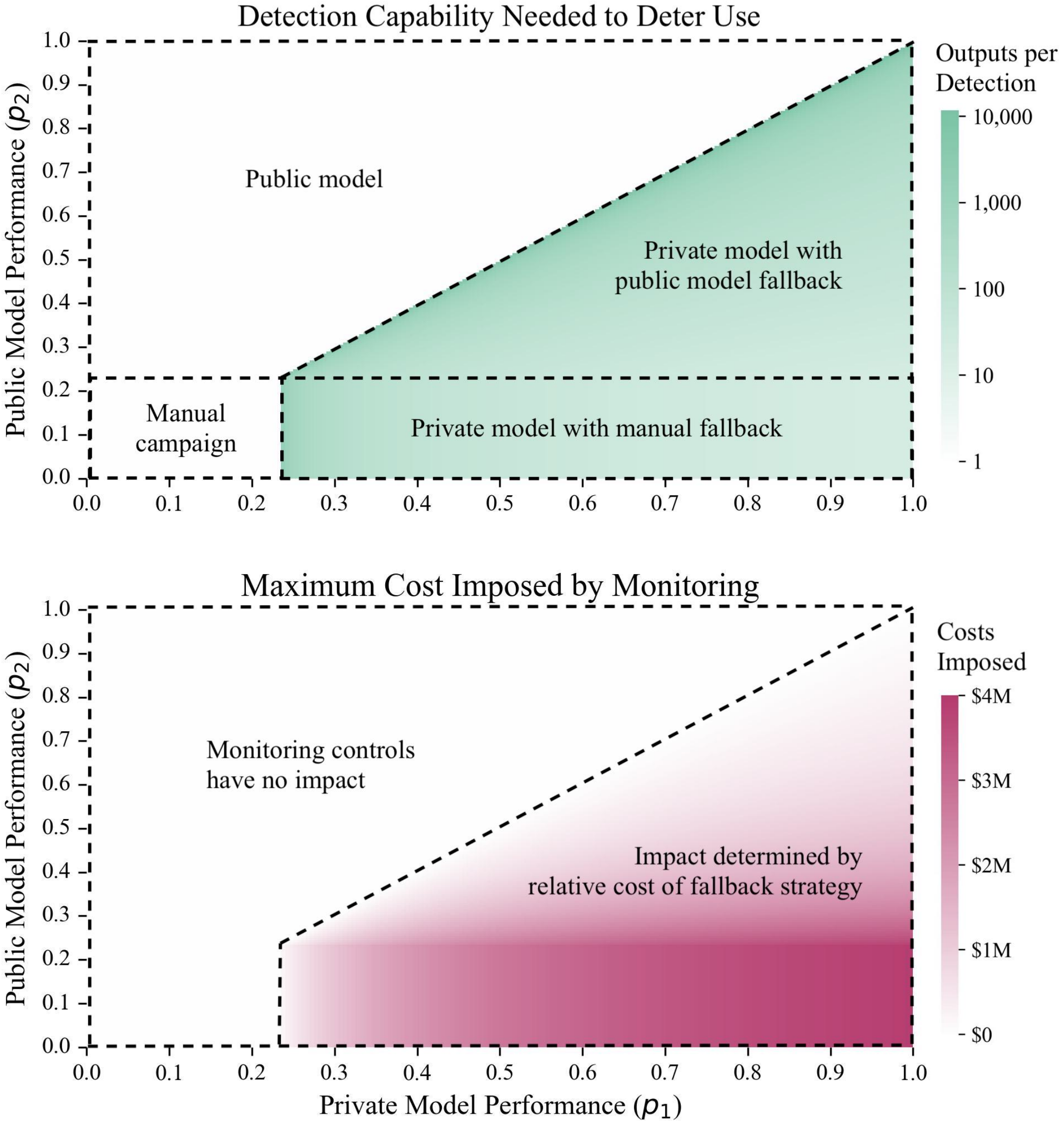

Figure 4: Optimal Strategies and Maximum Costs Imposed as a Function of p_1 and p_2

The lower right-hand rectangle of the bottom graph in Figure 4 shows similar information as Figure 3: with ideal detection capabilities, the maximum cost imposition of monitoring controls ranges from under $1,000,000 to roughly $3,500,000 over the course of a ten-million-tweet campaign, depending on the value of p_1. However, Figure 4 further shows that these cost impositions are dramatically reduced when operators can instead switch to alternative public models, even if those models perform less well. For instance, if both models perform with $p > 0.5$, but the best available public model consistently performs only 90% as well as the private model, then optimal detection capabilties (roughly, one detection per 500 outputs) will impose under $250,000 in additional costs.

6 Fixed Costs Associated with Running and Training Local Language Models

The previous discussion focuses entirely on variable costs associated with a manual campaign, the use of a public model, or the use of a private, monitored model. For simplicitly, I have treated the fixed costs associated with each type of campaign as negligible.[21] However, a propagandist need not treat the performance of an open source model as fixed; instead, they can choose to expend some additional up-front resources on finetuning the model to improve it. Let FR represent the feasibility region consisting of all points (FC, p_2) for which it is feasible to reach a given performance by expending FC in fixed costs.[22] Then the total expected costs facing the operator are given by:[23]

$$C(n) = \min \begin{pmatrix} \text{Manual} = \dfrac{nw}{L} \\ \text{API-Gated AI} = \dfrac{n}{p_1}\left(\dfrac{w}{\alpha L} + IC + P\lambda\right) \\ \text{Local AI} = \dfrac{n}{p_2}\left(\dfrac{w}{\alpha L} + IC\right) + FC, \quad (p_2, FC) \in FR \end{pmatrix} \tag{5}$$

The question of defining the feasibility region, that is, of articulating what types of capabilities are possible at various levels of investment, is both task-specific and requires advanced technical knowledge that goes well beyond the scope of this paper. Nonetheless, existing knowledge can be used to make some general estimates, as in the following scenarios:

- Suppose that ChatGPT-3.5 is capable of producing "usable" outputs for an operator at a rate of 0.85, but that ChatGPT-4 has a higher success rate of 1.0.[24] Because ChatGPT-4 requires a \$20 monthly fee to access, while ChatGPT-3.5 does not, we can treat the penalty for detection from ChatGPT-4 as \$20 higher than the penalty for detection from ChatGPT-3.5 (assuming that the propagandist will be detected at least once per month, and thus effectively pays this as a one-time signup fee after each detection). Plugging these values into the line for the API-gated AI in equation 5, using our Monte Carlo estimations for the other parameters, and rearranging, we can estimate that the propagandist will prefer to use ChatGPT-3.5 as long as $\lambda > 0.0009$, or as long as the propagandist can, on average, produce fewer than roughly 1,000 outputs before being detected.
- Suppose further that OpenAI has in fact implemented monitoring controls sufficient to detect malicious action at this rate. However, the propagandist can also reach performance on par with ChatGPT-3.5 by expending only \$600 to download and finetune an existing public model, i.e. the point $(\$600, 0.85) \in FR$.[25] Again plugging the relevant values into equation 5 and using Monte Carlo estimation for the remaining parameters, we can estimate that the propagandist will prefer to use the locally-run model if they anticipate using it for more than roughly 250,000 outputs.
- Finally, suppose that the propagandist cannot further improve any existing open source language models beyond this threshold with additional fine-tuning. However, the propagandist can choose to train a more advanced model of their own for \$4,600,000 which could perform as well as ChatGPT-4 (but without the

[21] A manual campaign theoretically does require upfront costs to acquire talent and create infrastructure. In fact, these fixed costs may dwarf the actual content generation costs, because maintaining a large infrastructure of accounts is often more costly than content production. But for this cost modeling exercise, a more realistic threat scenario is that of an already-existing propaganda outlet deciding whether or not to begin producing content using language models, a scenario under which the fixed costs for the manual campaign have already been paid. There may also be fixed costs associated with the initial account creation and familiarization with the technology involved in the use of a private, API-gated private model, but I treat these costs as negligble. Downloading and running a public model on local infrastructure likely carries the greatest fixed costs, but these can be incorporated into the following adjusted model; see footnote 22, below.

[22] I frame this as a question of expending fixed costs to improve a model through finetuning, but if there are fixed costs to running a model in the first place, this can be handled by defining FR such that any point in the set $\{FC, p_2 : FC = 0, p_2 > 0\} \notin FR$.

[23] Note that equation 5 can be expanded to include relevant comparisons of multiple API-gated models or multiple locally-running models. For instance, the feasibility region of an existing, relatively small model may not include high values of p which could be achieved were a propagandist to train a larger model from scratch, though such training might require much higher fixed costs. Similarly, multiple API-gated models may exist with different detection capabilities and performances.

[24] [19] used GPT-3 to generate articles making the same claims as a subset of known propaganda articles and compared their relative impact on readers' beliefs. Out of 18 articles generated using GPT-3, only two were deemed not relevant to the thesis statement intended by the researchers, for a success rate of 89%. Although this is a low threshold, so too is my threshold of "usable" outputs. In reality, due to additional safety measures implemented by OpenAI, ChatGPT-4 is likely to actually respond to malicious requests far less often than with perfect accuracy, but because a value of 1.0 represents the maximum benefit of the model, I use it here for illustrative purposes.

[25] In [58], researchers were able to achieve GPT-3.5-level performance for less than \$600 in finetuning expenses.

restrictions on access).[26] If the OpenAI models were the only ones available, training this model would be cost-effective if the propagandist planned to engage in influence operations requiring roughly 330 million outputs or more. But at that scale, using the $600-finetuned open source model is more cost effective than either OpenAI model, and compared to **that** alternative, the propagandist only finds it cost-effective to train a model from scratch if they intend to conduct campaigns larger than 430 million outputs in size.[27]

Although they require a lot of suppositions, these scenarios are useful for illustrating some general points: for very small campaigns, propagandists are likely to prefer using API-accessible models, even if those models have monitoring controls that impose significant costs. But given only moderate assumptions about the payoffs of finetuning small, lightweight models to perform propaganda-specific tasks, it very quickly becomes more cost-effective for operators to rely on such models. And even when those models still have relatively large limitations that necessitate continued and careful human curation of outputs, training a large language model from scratch is almost never economically worthwhile except at extremely large scales.

7 Sensitivity Analysis

The previous results include the following specific estimates:

1. **Threshold Performance**: A marginal output produced by a human-AI team is expected to become cheaper than a marginal output written by a human author whenever a language model is able to produce usable outputs at a rate higher than 0.23.

2. **Maximum Savings**: Over the course of a 10-million-tweet campaign, with a language model that produces usable outputs at a rate of 75%, a propagandist could expect to save $3 million in content generation costs, on average (equivalent to a 70% reduction in costs, assuming no fixed costs to using the model and no monitoring controls in place on the model).

3. **Optimal Detection Rate (API Only)**: If an operator can access an API-gated model that produces usable outputs at a rate of 75%, and if their only fallback in response to costly monitoring controls is to resort to human authorship, then monitoring controls that deny the full potential savings from the use of AI require a minimum detection capability of roughly one detection per 40 outputs.

4. **Maximum Cost Imposition (Public Option)**: However, if an open source but slightly worse-performing model (say, one that produces usable outputs at a rate of 70%) exists, then the maximum cost imposition generated by monitoring controls is $720,000.

5. **Minimum Viable Size (Finetuning vs. API)**: If a propagandist can finetune an existing open source model for $600 to produce usable outputs at a rate of 85%, and if a similarly capable API-gated model exists that is monitored with a rate of one detection per every 1,000 outputs, then the finetuned model is preferred for any campaign larger than roughly 130,000 outputs.

6. **Minimum Viable Size (Training vs. Finetuning)**: However, if reaching a performance reliability of 100% requires training an LLM from scratch at roughly the cost of GPT-3's original training run ($4.6 million), the training from scratch is only cost effective for campaigns larger than roughly 430 million outputs.

These point estimates require a number of parameters to be manually specified, primarily models' performance rates, detection capabilities, and fixed costs (at least when these parameters themselves are not the object of analysis). However, the estimates also rely on Monte Carlo parameter estimation for five key variables: α, w, L, IC, and P. Table 1 provides summaries for the ranges over which values for each variable were (uniformly) sampled, and reiterates the general source(s) from which each of these ranges were extrapolated.

[26] [23] estimates the total cost of training GPT-3 at $4,600,000. For this scenario, I imagine that the performance of GPT-4 could be replicated with a GPT-3-sized model plus finetuning, but that it could not be replicated with a smaller model.

[27] Note that, as per the final paragraph of Section 2, if the propagandist were able to produce a model that could operate **fully** autonomously, this expenditure would become cost-effective at campaigns of only 10.7 million outputs. However, this would require no human involvement even for the identification of posts to comment on and the writing of prompts to generate those posts from the AI model.

Table 1: Sampling Ranges for Key Parameters

Parameter	Lower Bound	Upper Bound	Midpoint	Justification
α	2	10	6	Observed Values from Code Generation Tasks
w	\$1.41	\$9.53	\$5.47	Historical IRA Job Postings and Operations
L	5	25	15	Historical IRA Job Postings and Operations
IC	\$0.0006	\$0.024	\$0.01 (approx.)	API Fees for Existing Models
P	$0.5w$	$2w$	$1.25w$	Optimistic Range of Impact

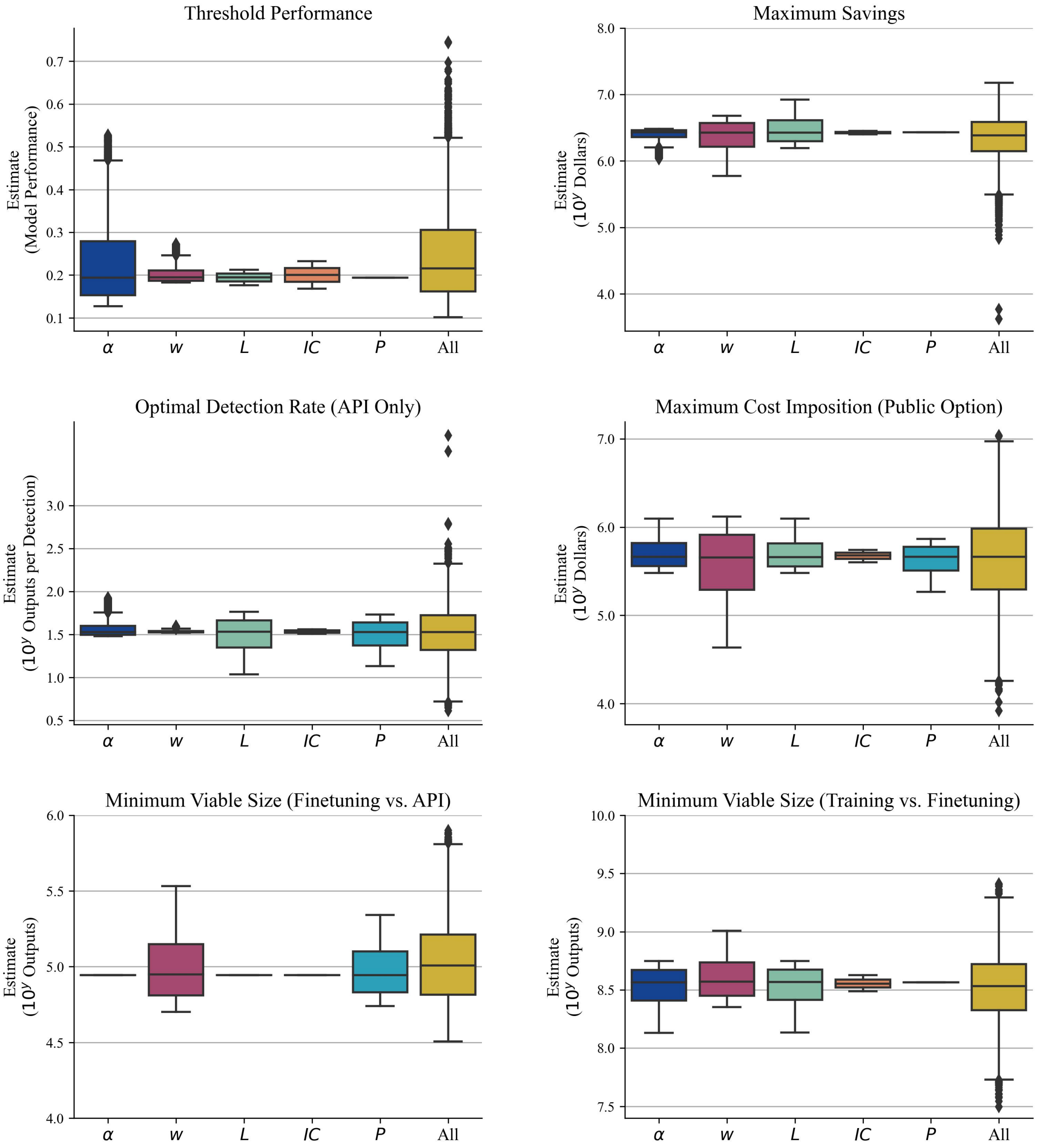

Figure 5: Contributions from Uncertainty in Key Parameters to Variation in Overall Results

While all of these sampling ranges span large regions of uncertainty, uncertainty in some parameter estimates more strongly drive variation in point estimates for the above numerical results than uncertainty in others. Figure 5 attempts to visualize the way that uncertainty in each parameter contributes to variation in estimates for the four numerical results listed above. Each boxplot represents 10,000 estimates of the parameter of interest where only the parameter on the x-axis is allowed to vary; all other parameters are held constant at their midpoint value. The final boxplot on the right-hand side of each subplot shows the variation in the parameter estimate when all parameters listed along the x-axis are allowed to vary on the ranges shown in Table 1.

The top-left figure, for instance, shows that the vast majority of variance in the predicted threshold performance at which it becomes cost effective to use a language model depends on α, and not on the inference costs, wages, or labor productivity of propagandists. This is unsurprising, as equation 1—which states when it is profitable to use a human-machine team at all—loosely approximates to an inverse relationship between p and α (when IC takes small values). By contrast, other parameters of interest are more heavily determined by values of w, L, and—for estimates where monitoring controls on an API-gated model are relevant—P. No parameter estimate heavily depends on variation in IC, which further justifies treating the inference costs of different models as equivalent (see footnote 18).

Several other points are worth making regarding Figure 5. Interestingly, for example, the estimates for the potential savings that unrestricted access to a reasonably reliable language model could generate (Maximum Savings) span roughly two orders of magnitude. But the estimates for the maximum potential cost imposition of monitoring controls when an slightly-less-reliable alternate public model exists (Cost Imposition [Public Option]) span a wider three orders of magnitude. There is also relatively large uncertainty regarding the detection capability needed to fully deter propagandists from using an API-gated model (when public options do not exist). There is less variation, by contrast, in the estimates regarding the scale of operation at which a finetuned model becomes more cost effective than a similarly-performing API-gated model (and such variation exclusively depends on the penalty imposed by each detection, which is itself a function of wages).

8 Discussion

The preceding analysis suggests that, under a relatively wide range of potential scenarios, the use of language models to produce misinformation content is highly cost effective, relative to the use of purely manual content generation. While it is not surprising that a fully automated campaign would be cheaper than paying humans to write content for influence operations, these models suggest that the use of even relatively unreliable models can substantially reduce propagandists' costs via human-machine teams, as long as models produce "usable" outputs more often than one in four times. With human-machine teams, labor costs still dominate an operation's overall content generation costs, but savings can quickly approach the millions of dollars.

Before concluding, it is worth discussing a few general comments on the implications and limitations of this work.

8.1 Model limitations

This model of influence operations is limited in a number of key ways. First, and most notably, parameter estimates regarding worker productivity and wages in existing influence operations are based on a small number of investigative reports or job postings, almost exclusively in the context of Russian influence operations. These figures may or may not generalize to other propaganda operations, but an absence of public data about the organization and economic structures of propaganda campaigns makes further precision difficult. Additionally, while some research has begun to emerge regarding the impact of LLM usage on worker productivity, [7, 14, 9, 56], there is still large uncertainty regarding how effective disinformation operators will be at incorporating LLMs into their workflows. Future economic research in other domains may significantly help to narrow the uncertainties in this model.

Relatedly, the model analyzes the economics of using LLMs for discrete tasks insofar as it uses a single value—p—to describe a model's capability. But p is not meant to be a description of a model's abstract capabiltiies, but rather its reliability at producing usable outputs on a specific task. For instance, a model may perform reliably enough to save money on the task of tweet generation, but struggle more with longer-form content, making it cost ineffective for use on the task of fake news article generation. To **some** extent, it may be reasonable to interpret a finding that "use of such and such a model becomes cost effective given certain assumptions at x outputs" as meaning that the model becomes cost effective when used across multiple tasks to produce any type of written content as long as x tweets. But such an inference relies on the assumption that performance across tasks is relatively consistent, which may or may not be true. In other words, while the models used here can give a rough sense of the scale at which certain content production strategies become cost saving, they do not fully capture the multiplicity of content generation tasks for which actual propagandists would be likely to use LLMs.

This model is also strictly focused on the cost savings associated with producing content at scale, and not with quality improvements that LLMs could offer propagandists. But major quality improvements may be possible, providing an additional incentive to makes use of LLMs in propaganda campaigns. The use of copypasta, stilted language from non-native speakers, or transliterations of idioms that do not make sense in a propagandist's target language have often provided important clues regarding inauthentic behavior. [12, 21, 55] These errors in human cultural awareness and translation ability are more prevalent in influence operations conducted by some countries than others, and countries whose propagandists frequently make such errors may be forced to prioritize volume in output instead of taking time to carefully craft believable—and more persuasive—personas for their fake accounts. [20, 35] Language models, by contrast, are unlikely to make such easily-noticeable mistakes, though they may still struggle to effectively mimic discursive norms common in fine-grained target populations. [6, 8, 52] Other research has noted that content produced from GPT-3 can change readers' opinions on sensitive political issues, and can do so even better than existing examples of propaganda news articles with only light copyediting. [19] These improvements in quality may permit propagandists to meaningfully alter the strategies that they employ in influence operations. [18]

8.2 Factors other than cost may disincentivize the use of LLMs in influence operations

Even if it is strictly cost-effective for malicious actors to use LLMs to produce disinformation content, organizational, bureaucratic, or cultural barriers may cause propaganda outlets to nonetheless avoid doing so. Propaganda outlets can take a number of forms. [15] describes the proliferation of private "disinformation-for-hire" firms that are contracted to generate and promote disinformation, whereas [25] argues that large quantities of Chinese-origin propaganda on social media is produced by a diffuse set of government bureaucrats who are paid per output to produce propaganda content but without any centralized direction or oversight. Propaganda outlets organized around the first model will likely have much stronger incentives to adopt cost-saving technologies than those operating on the second model. In the extreme, we can even speculate that bureaucratic structures which reward departments on the basis of personnel size may actively disincentivize the adoption of LLMs for propaganda purposes.[28]

More generally, it is unclear to what extent propagandists are optimizers or satisficers. The development of "deepfake" technology over the 2010s led some analysts to speculate that Russia would unleash a "wave" of deepfaked disinformation against the West. [38] However, despite some high-profile examples [1], deepfaked images and videos remained an apparently minor component of Russian influence operations for relatively long after the technology to produce them existed. Only in recent months, with the rise of text-to-image generative AI systems, have AI-generated images become more commonly observed as a tool for disinformation (though not necessarily as a tool of Russian disinformation specifically). [49, 42] This potentially suggests that technical barriers to adoption can meaningfully deter propgandists from using new technologies, and that improvements in user-friendliness affect propagandists' decision-making more than improvements in underlying capabilities.

Similarly, although the use of LLMs to produce disinformation was largely speculative until recently, recent months have seen networks of Twitter accounts posting tweets containing ChatGPT's default refusal to complete a user request response, likely suggesting attempts to use the model to generate content to post on social media. [41] While the models presented here suggest that for even relatively small campaigns, a propagandist's cost-optimal solution is to finetune an open source model, propagandists may instead prefer to rely on solutions with lower adoption costs even when doing so is economically irrational.

8.3 Nation-states do not have strong incentives to secretly train LLMs for influence operations

The maximum length of a campaign evaluated in the preceding sections was 10 million tweets. This volume of coordinated inauthentic activity on Twitter between October 2018 and December 2021 was exceeded by only three countries (see Section 3). However, this is true only when considering (1) publicly attributed activity that was (2) posted to and removed by Twitter (3) over a three-year period with significant gaps in reporting. It seems reasonably likely that for a small but significant number of nation-state actors, the amount of content generated for use in influence operations over the near- to medium-term could substantially exceed the equivalent of 10 million tweets. In fact, [25] estimates that the Chiense government fabricates and posts "about 448 million social media comments a year."

However, even at this scale, the value of training an LLM from scratch to produce disinformation content—as opposed to simply finetuning an existing open source model—is dubious, even if the finetuned model is not as capable (see

[28] It is not clear whether any major propaganda outlets face this set of incentives. However, one common issue facing propagandists is that it is remarkably difficult to evaluate the impact of disinformation on actual political attitudes and behaviors, with some research indicating that the concrete effects of exposure to influence operations is relatively small. [13] The difficulty of evaluating the political import of specific operations, combined with the insulation from cost-cutting pressures that exists for in-house propaganda authors (contrasted with specialized firms), could largely nullify the incentives to adopt LLMs for content generation.

Section 6). If the best attainable performance of any open source model, even after careful finetuning, was still very low for a given task, it **might** become economically viable to train an LLM from scratch. But the propagandist must not only believe this to be true of existing open source models, but also of any future models that may be released between the time the operator begins training their model and the time they generate enough posts to fully recoup their expenses. Given the rapid rate of both public model release and the tendency of access to privately-held models to become easier and less restricted over time, this is unlikely to be a reasonable bet.[29]

It is possible that the use of LLMs may themselves enable much larger campaigns, such that although training a model from scratch would be a poor economic decision under **current** scales of operation, doing so would enable much larger scales of operation that **do** justify such an investment. But there are checks on the scale at which propagandists can operate that go beyond the costs of content generation, including the difficulty of maintaining large networks of inauthentic accounts without being detected by platforms. [17, 51] It seems reasonably likely, then, that even nation-states may find it difficult to justify secretive large-scale training runs of LLMs intended primarily for use in influence operations.

8.4 The comparative value of technical mitigations against LLM misuse

There are three broad ways in which LLM developers can reduce the likelihood of their models being abused: they can train or finetune the model itself in ways that reduce its propensity to comply with malicious user requests (thereby reducing p), they can invest in capabilities to detect misuse or impose greater penalties on identified malicious actors (thereby increasing $P\lambda$),[30] or they can embed watermarks into model outputs or pursue other strategies that increase the potential for detection of synthetic content online.[31] The model and analysis presented here indicates that all three strategies can be valuable, though in different ways.

Model alignment efforts that reduce p and monitoring controls that increase $P\lambda$ are primarily useful when rival open source models do not exist, or if propagandists are particularly "sticky" and unlikely to switch to such rival models. Model alignment efforts can also be pursued by groups who develop and release open source models, though to date, this is less common than among businesses that seek to monetize their models. However, monitoring controls are not in principle applicable to open source models.[32]

The development of watermarks for LLMs is an active area of research. Existing proposals for technical methods of watermarking LLMs, however, are "shallow" in the sense that they are added on top of a pretrained LLM and can easily be removed by a user or eliminated via finetuning. [26] However, a growing number of researchers are exploring the feasibility of "deep" watermarks or other methods that persist and allow for attribution even after finetuning. [2, 37, 46] Whether or not such deep watermarks will prove to be feasible is an open question—but if they are, and if propagandists looking to use LLMs for content generation primarily rely on finetuned versions of open source models, then embedding such watermarks into open source models may be a valuable intervention.

[29] Note also that, even if the Chinese government produces roughly 448 million inauthentic social media posts per year, this volume of content is likely not produced by a single propaganda agency that could pay the initial fixed costs of model training and then recoup their expenses over time; rather, it is produced (at least in part) by a diffuse set of bureaucrats, no one member of which stands to gain from paying large upfront costs for the sake of increasing their individual efficiency at generating misinformation. [25]

[30] "Increasing penalties" here can mean anything that imposes additional friction upon propagandists once identified. For instance, requiring a CAPTCHA in addition to an email address for users signing up for model access imposes additional costs, though not very large ones.

[31] The model presented here does not readily include a way for analyzing this strategy. Watermarks reduce the value of LLM outputs, but they do not make it more costly to generate the outputs, meaning that a strict cost comparison does not capture the relevant differences.

[32] Note that even if these interventions do not carry major benefits from a security perspective, they may still be valuable from a safety perspective. In addition, if propagandists are satisficers, it is possible that a failed attempt to make use of an API-gated model (whether due to the model's refusal to produce the desired outputs or a quick detection) may dissuade them from seriously pursuing the use of LLMs by other means as well.

A Supplemental Equations

The model presented in this paper requires only algebraic manipulation, primarily of equation 5, in order to calculate any of the final variables of interest discussed in Section 7. However, some of this algebraic manipulation can be tedious, so I reproduce here some useful solutions for various parameters of interest.

First, let $\hat{p}_{1H}$ represent the threshold performance value at which the use of the private model is preferred to reliance on a manual campaign (assuming the detection rate λ is already fixed). Then this value is given by:

$$\hat{p}_{1H} = \frac{1}{\alpha} + \frac{L}{w}\left(IC + P\lambda\right) \tag{6}$$

Let $\hat{p}_{1AI}$ represent the threshold performance value at which the use of the private model is preferred to an alternative open source option. This value is given by:

$$\hat{p}_{1AI} = p_2 + \frac{p_2 P\lambda - \frac{FC}{n}}{\frac{w}{\alpha L} + IC} \tag{7}$$

Note that, if $p_1 = p_2$, the propagandist is only indifferent between the API-gated and open source models if $p_2 P\lambda = \frac{FC}{n}$. In other words, the API-gated model can perform worse than the open source model and face meaningful monitoring risks, and yet still be preferred if $\frac{FC}{n}$ is sufficiently large.

If, alternatively, values of p_1 are already known, we may instead want to calculate the minimum detection capability at which a propagandist is deterred from using the API-gated model. If the propagandist's fallback to the use of the API-gated model is a manual campaign, then the minimum deterrant detection capability $\hat{\lambda}_H$ is given by:

$$\hat{\lambda}_H = \frac{1}{P}\left(\frac{wp_1}{L} - \frac{w}{\alpha L} - IC\right) \tag{8}$$

Alternatively, if the public model is sufficiently well-performing that the propagandist will use it instead as a fallback, then the minimum deterrant detection capability $\hat{\lambda}_{AI}$ is instead given by:

$$\hat{\lambda}_{AI} = \frac{1}{P}\left(\frac{w}{\alpha L} + IC\right)\left(\frac{p_1 - p_2}{p_2}\right) + \frac{p_1 FC}{nP} \tag{9}$$

Note that in the two preceding equations, a detection penalty of \$0 causes $\hat{\lambda}$ to be undefined, because no detection capability could possibly deter malicious use if the detection does not itself impose some form of penalty. In addition, note that the equation 9 is the general case of equation 4, where fixed costs associated with running an open source model are no longer assumed to be \$0.

Finally, let equation 5 represent the three choices for operation facing the propagandist, but let it also be possible for the propagandist to spend FC (or ΔFC more than was spent for the use of the current open source model option) to create or finetune a new model with target capability $\hat{p}$. Then, for each of the three campaign styles, we can set $\frac{n}{\hat{p}}\left(\frac{w}{\alpha L} + IC\right) + FC$ equal to the corresponding portion of equation 4 and solve for n to calculate the minimum campaign size at which expending FC becomes cost effective relative to an existing choice of model (including manual authorship as a "choice" of model). The overall minimum viable scale for a model where $(FC, \hat{p}) \in FC$ is then the maximum solution of n across all alternative model choices (because the minimum viable scale is the scale at which a model becomes cost effective relative to the next-most-cost-effective option). This is given by:

$$\hat{n} = \max \begin{pmatrix} \text{Manual} = \dfrac{\hat{p}FC}{\frac{\hat{p}w}{L} - \left(\frac{w}{\alpha L} + IC\right)} \\ \text{API-Gated AI} = \dfrac{(\hat{p} \cdot p_1)FC}{\hat{p}P\lambda + (\hat{p} - p_1)\left(\frac{w}{\alpha L} + IC\right)} \\ \text{Local AI} = \dfrac{(\hat{p} \cdot p_2)\Delta FC}{(\hat{p} - p_2)\left(\frac{w}{\alpha L} + IC\right)} \end{pmatrix} \tag{10}$$

References

[1] Allyn, Bobby. "Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warm." *NPR*. March 16, 2022. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[2] Anonymous. "Towards Robust Model Watermark via Reducing Parametric Vulnerability." *Open Review preprint*. Accessed June 7, 2023. https://openreview.net/pdf?id=wysXxmukfCA.

[3] Baele, Stephane. "Artificial Intelligence and Extremism: The Threat of Language Models for Propaganda Purposes." Centre for Research and Evidence on Security Threats: October 25, 2022. https://crestresearch.ac.uk/resources/artificial-intelligence-and-extremism-the-threat-of-language-models/.

[4] Bagdasaryan, Eugene and Vitaly Shmatikov. "Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures." *arXiv [cs.CR]*. December 9, 2021. 10.48550/arXiv.2112.05224.

[5] Böswald, Lena-Maria and Beatriz Almeida Saab. "What a Pixel Can Tell: Text-to-Image Generation and its Disinformation Potential." Democracy Reporting International: September 2022. https://democracyreporting.s3.eu-central-1.amazonaws.com/images/6331fc834bcd1.pdf.

[6] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language Models are Few-Shot Learners." *arXiv [cs.CL]*. May 28, 2020. 10.48550/arXiv.2005.14165: 35.

[7] Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. "Generative AI at Work." NBER Workiing Paper No. 31161. https://www.nber.org/system/files/working_papers/w31161/w31161.pdf.

[8] Buchanan, Ben, Andrew Lohn, Micah Musser, and Katerina Sedova. "Truth, Lies, and Automation: How Language Models Could Change Disinformation." Center for Security and Emerging Technology: May 2021. 10.51593/2021CA003.

[9] Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliviera Pinto, Jared Kaplan, and Harri Edwards et al. "Evaluating Large Language Models Trained on Code." *arXiv [cs.LG]*. July 7, 2021. https://arxiv.org/abs/2107.03374.

[10] Cohere. "Pricing." Accessed June 7, 2023. https://cohere.ai/pricing.

[11] Cohere, OpenAI and AI21 Labs. "Best Practices for Deploying Language Models." *OpenAI Blog*. June 2, 2022. https://openai.com/blog/best-practices-for-deploying-language-models/.

[12] DiResta, Renée, Josh A. Goldstein, Carly Miller and Harvey Wang. "One Topic, Two Networks: Evaluating Two Chinese Influence Operations on Twitter Related to Xinjiang." Stanford Internet Observatory & Stanford Cyber Policy Center: December 2, 2021. https://cyber.fsi.stanford.edu/io/publication/one-topic-two-networks.

[13] Eady, Gregory, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. "Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior." *Nature Communications* 14 (2023): 10.1038/s41467-022-35576-9.

[14] Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." *arXiv [econ.GN]*. March 17, 2023. https://arxiv.org/abs/2303.10130.

[15] Fisher, Max. "Disinformation for Hire, a Shadow Industry, Is Quietly Booming." *New York Times*. July 25, 2021. https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html.

[16] Gilbert, David. "Inside Cyber Front Z, the 'People's Movement' Spreading Russian Propaganda." *Vice News*. April 4, 2022. https://www.vice.com/en/article/7kbjny/russia-cyber-front-z-telegram.

[17] Goldstein, Josh, Girish Sastry, Sarah Kreps, and J.D. Maddox. "Large Language Models and the Future of Disinformation." YouTube. November 15, 2022. Panel discussion, 58:57. https://www.youtube.com/watch?v=ev07nlTBZ3Q.

[18] Goldstein, Josh, Girish Sastry, Micah Musser, Matthew Gentzel, Renée DiResta, and Katerina Sedova. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." *arXiv [cs.CY]*. January 11, 2022. https://cdn.openai.com/papers/forecasting-misuse.pdf.

[19] Goldstein, Josh, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. "Can AI Write Persuasive Propaganda?" *SocArXiv*. February 21, 2023. 10.31235/osf.io/fp87b.

[20] Goldstein, Josh and Renée DiResta. "China's Fake Twitter Accounts Are Tweeting Into the Void." *Foreign Policy*. December 15, 2021. https://foreignpolicy.com/2021/12/15/china-twitter-trolls-ccp-influence-operations-astroturfing/.

[21] Graphika, Stanford Internet Observatory, and Stanford Cyber Policy Center. "Unheard Voice: Evaluating five years of pro-Western covert influence operations." August 24, 2022. https://cyber.fsi.stanford.edu/io/news/sio-aug-22-takedowns.

[22] Hwang, Tim. "The Microeconomics of Disinformation." YouTube. October 12, 2022. Lecture, 1:00:05. https://www.youtube.com/watch?v=JJZObKWG8ok.

[23] Isahq, Rana. "How much did GPT-3 cost?" *PC Guide*. March 22, 2023. https://www.pcguide.com/apps/gpt-3-cost/.

[24] Kalliamvakou, Eirini. "Research: quantifying GitHub Copilot's impact on developer productivity and happiness." *GitHub Blog*. September 7, 2022. https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/.

[25] King, Gary, Jennifer Pan, and Margaret E. Roberts. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." *American Political Science Review* 111, no. 3 (2017): 484–501. https://doi.org/10.1017/S0003055417000144.

[26] Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. "A Watermark for Large Language Models." *arXiv [cs.LG]*. January 24, 2023. https://arxiv.org/abs/2301.10226.

[27] Klochkova, Ksenia. "'You don't believe these are real reviews, do you?' How Fontanka looked at the front line of Cyber Front Z." *Fontanka*. March 21, 2022. https://archive.ph/TB4Xw.

[28] Korinek, Anton. "Language Models and Cognitive Automation for Economic Research." NBER Working Paper 30957. https://www.nber.org/system/files/working_papers/w30957/w30957.pdf.

[29] Kreps, Sarah, R. Miles McCain and Miles Brundage. "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation." *Journal of Experimental Political Science* 9, no. 1 (Spring 2022). https://doi.org/10.1017/XPS.2020.37.

[30] Lazar, Seth. "Ethics for Generative Agents." *Normative Philosophy of Computing* [lecture notes]. February 23, 2023. https://write.as/sethlazar/genb.

[31] Leahy, Connor. "Why Release a Large Language Model?" *EleutherAI Blog*. June 2, 2021. https://blog.eleuther.ai/why-release-a-large-language-model/.

[32] Lohn, Andrew and Krystal Jackson. "Will AI Make Cyber Swords or Shields: A few mathematical models of technological progress." *arXiv [cs.CR]*. July 27, 2022. https://arxiv.org/abs/2207.13825.

[33] Lytvynenko, Jane. "Here Are Some Job Ads For The Russian Troll Factory." *BuzzFeed News*. February 22, 2018. https://www.buzzfeednews.com/article/janelytvynenko/job-ads-for-russian-troll-factory.

[34] Lyu, Siwei. "Deepfakes and the New AI-Generated Fake Media Creation-Detection Arms Race." *Scientific American*. July 20, 2020. https://www.scientificamerican.com/article/detecting-deepfakes1/.

[35] Mandiant Intelligence. "Pro-PRC DRAGONBRIDGE Influence Campaign Leverages New TTPs to Aggressively Target U.S. Interests, Including Midterm Elections." *Mandiant*. October 26, 2022. https://www.mandiant.com/resources/blog/prc-dragonbridge-influence-elections.

[36] McGuffie, Kris and Alex Newhouse. "The Radicalization Risks Posed by GPT-3 and Adavanced Neural Language Models." Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studes: September 2020. https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf.

[37] Merkhofer, Elizabeth, Deepesh Chaudhari, Hyrum S. Anderson, Keith Manville, Lily Wong, and João Gante. "Machine Learning Model Attribution Challenge." *arXiv [cs.LG]*. February 13, 2023. https://arxiv.org/abs/2302.06716.

[38] Meserole, Chris and Alina Polyakova. "The West is ill-prepared for the wave of 'deep fakes' that artificial intelligence could unleash." *Brookings Institution*. May 25, 2018. https://www.brookings.edu/blog/order-from-chaos/2018/05/25/the-west-is-ill-prepared-for-the-wave-of-deep-fakes-that-artificial-intelligence-could-unleash/.

[39] Murali, Vijayaraghavan, Chandra Maddila, Imad Ahmad, Michael Bolin, Daniel Cheng, Negar Ghorbani, Renuka Fernandez, and Nachiappan Nagappan. "CodeCompose: A Large-Scale Industrial Deployment of AI-assisted Code Authoring." *arXiv [cs.SE]*. May 20, 2023. https://arxiv.org/abs/2305.12050.

[40] Narayanan, Arvind and Sayash Kapoor. "The LLaMA is out of the bag. Should we expect a tidal wave of disinformation?" *AI Snake Oil*. March 6, 2023. https://aisnakeoil.substack.com/p/the-llama-is-out-of-the-bag-should.

[41] Norteño, Conspirador (@conspirator0). Twitter Post. April 16, 2023, 2:40PM. https://archive.is/UoMM9.

[42] Norton, Tom. "Face Check: Photo of Putin on His Knees in Front of China's Xi." *Newsweek*. March 22, 2023. https://www.newsweek.com/fact-check-photo-putin-his-knees-front-chinas-xi-1789498.

[43] OpenAI. "DALL·E Now Available Without Waitlist." September 28, 2022. https://openai.com/blog/dall-e-now-available-without-waitlist/.

[44] OpenAI. "OpenAI's API Now Available with No Waitlist." November 18, 2021. https://openai.com/blog/api-no-waitlist/.

[45] OpenAI. "Pricing." Accessed June 7, 2023. https://openai.com/api/pricing/.

[46] "Overview." *Machine Learning Model Attribution Challenge*. Accessed June 7, 2023. https://mlmac.io/.

[47] Patel, Andrew and Jason Sattler. "Creatively malicious prompt engineering." *WithSecure Intelligence*. January 11, 2023. https://labs.withsecure.com/publications/creatively-malicious-prompt-engineering.

[48] Radford, Alec, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. "Better Language Models and Their Implications." *OpenAI Blog*. February 14, 2019. https://openai.com/blog/better-language-models/.

[49] Sarkar, Alisha Rahaman. "Chilling AI deepfakes purporting to show Trump arrest take over Twitter." *The Independent*. March 24, 2023. https://www.independent.co.uk/news/world/americas/us-politics/trump-deepfake-arrest-twitter-ai-b2307470.html.

[50] Seddon, Max. "Documents Show How Russia's Troll Army Hit America." *Buzzfeed News*. June 2, 2014. https://www.buzzfeednews.com/article/maxseddon/documents-show-how-russias-troll-army-hit-america.

[51] Sedova, Katerina. "AI and the Future of Disinformation Campaigns, Part 1: The RICHDATA Framework." Center for Security and Emerging Technology: December 2021. 10.51593/2021CA005.

[52] Sedova, Katerina. "AI and the Future of Disinformation Campaigns, Part 2: A Threat Model." Center for Security and Emerging Technology: December 2021. 10.51593/2021CA011.

[53] Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." *arXiv [cs.CY]*. February 5, 2023. https://arxiv.org/abs/2302.04844.

[54] Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford et al. "Release Strategies and the Social Impacts of Language Models." *arXiv [cs.CL]*. August 24, 2019. 10.48550/arXiv.1908.09203: 6–8.

[55] Strick, Benjamin. "Twitter Analysis: Identifying a Pro-BJP Copypasta Influence Operation in India." *<Ben>*. January 11, 2021. https://benjaminstrick.com/twitter-analysis-identifying-a-pro-bjp-influence-operation-in-india/.

[56] Tabachnyk, Maxim and Stoyan Nikolov. "ML-Enahnced Code Completion Improves Developer Productivity." *Google Research Blog*. July 26, 2022. https://ai.googleblog.com/2022/07/ml-enhanced-code-completion-improves.html.

[57] Tamkin, Alex, Miles Brundage, Jack Clark, and Deep Ganguli. "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models." *arXiv [cs.CL]*. February 4, 2021. 10.48550/arXiv.2102.02503.

[58] Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. "Alpaca: A Strong, Replicable Instruction-Following Model." *Stanford University Center for Research on Foundation Models*. March 13, 2023. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[59] Tiku, Nitasha. "AI can now create any image in seconds, bringing wonder and danger." *Washington Post*. September 28, 2022. https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/.

[60] Twitter Transparency. "Information Operations." Accessed July 6, 2022. https://transparency.twitter.com/en/reports/information-operations.html.

[61] Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng et al. "Ethical and Social Risks of Harm from Language Models." *arXiv [cs.CL]*. December 8, 2021. 10.48550/arXiv.2112.04359.