



Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions

Max Weiss

Highlights

- Publicly available artificial intelligence methods can generate an enormous volume of original, human speech-like topical text (“Deepfake Text”) that is not based on conventional search-and-replace patterns
- I created a computer program (a bot) that generated and submitted 1,001 deepfake comments regarding a Medicaid reform waiver to a federal public comment website, stopping submission when these comments comprised more than half of all submitted comments. I then formally withdrew the bot comments
- When humans were asked to classify a subset of the deepfake comments as human or bot submissions, the results were no better than would have been gotten by random guessing
- Federal public comment websites currently are unable to detect Deepfake Text once submitted, but technological reforms (e.g., CAPTCHAs) can be implemented to help prevent massive numbers of submissions by bots

INCORRECT

0%

Was the following comment created by a bot or a human?

I am writing to you with regard to Idaho's proposed Medicaid waiver which has problems as it is currently written. Many Idahoans depend on Medicaid when they are sick and need help. Implementing the proposed waiver would mean taking away health care when people are most vulnerable. If someone has low income and becomes ill and cannot work that is not the time to take away their coverage.

Bot

X Human

Example Deepfake Text generated by the bot that all survey respondents thought was from a human.

Abstract

The federal comment period is an important way that federal agencies incorporate public input into policy decisions. Now that comments are accepted online, public comment periods are vulnerable to attacks at Internet scale. For example, in 2017, more than 21 million (96% of the 22 million) public comments submitted regarding the FCC's proposal to repeal net neutrality were discernible as being generated using search-and-replace techniques [1]. Publicly available artificial intelligence methods can now generate "Deepfake Text," computer-generated text that closely mimics original human speech. In this study, I tested whether federal comment processes are vulnerable to automated, unique deepfake submissions that may be indistinguishable from human submissions. I created an autonomous computer program (a bot) that successfully generated and submitted a high volume of human-like comments during October 26-30, 2019 to the federal public comment website for the Section 1115 Idaho Medicaid Reform Waiver.

Results summary: The bot generated and submitted 1,001 deepfake comments to the public comment website at Medicaid.gov over a period of four days. These comments comprised 55.3% (1,001 out of 1,810) of the total public comments submitted. Comments generated by the bot were often highly relevant to the Idaho Medicaid waiver application, including discussion of the proposed waiver's consequences on coverage numbers, its impact on government costs, unnecessary administrative burdens, and relevant personal experience. Finally, in order to test whether humans can distinguish deepfake comments from other comments submitted, I conducted a survey of 108 respondents on Amazon's Mechanical Turk. Survey respondents, who were trained and assessed through exercises in which they distinguished more obvious bot versus human comments, were only able to correctly classify the submitted deepfake comments half (49.63%) of the time, which is comparable to the expected result of random guesses or coin flips. This study demonstrates that federal public comment websites are highly vulnerable to massive submissions of deepfake comments from bots and suggests that technological remedies (e.g., CAPTCHAs) should be used to limit the potential of abuse.

Introduction

From April to October of 2017, the Federal Communications Commission (FCC) received a record-breaking 22 million submissions on its public comment website, offering public input on the regulatory proposal to repeal net neutrality protections under Title II of the Communications Act [1, 2]. Net neutrality refers to requirements that Internet service providers provide access to all Internet content without favoring or blocking particular websites [3]. Federal law requires comments from potentially affected individuals, businesses, and organizations to be taken into account by federal agencies in decision-making, though the specific extent of public comment consideration by any given agency may be unclear [2]. The 22 million submitted comments about net neutrality included about 5

million comments that supported net neutrality and about 17 million that supported its repeal. In a December 2017 decision, the FCC voted 3-2 to repeal net neutrality [4].

On its face, the FCC comment period represented an effective exercise in democratic accountability: A federal agency asked for public input, the agency received a large number of public comments, and the agency seemingly took the public comments into account to make its final decision.

However, soon after the end of the public comment period, researchers alleged that hundreds of thousands of the comments were submitted under fake names, some stolen and others completely fabricated [2]. Comments came from email addresses, street addresses, and postal codes stolen from unwitting victims, constituting countless instances of identity theft [5, 6].

Subsequent text analysis found that only 800,000 (less than 4%) of the comments submitted were likely to be unique and authentic [1, 7]. Computer programs that automated Internet tasks (bots) using simple search-and-replace techniques were responsible for generating and submitting the overwhelming majority of the 22 million comments, and those comments were lopsided, being part of coordinated campaigns to support net neutrality repeal. Twenty comment-duplication campaigns alone accounted for 17 million of the 22 million comments [1].

Most of the fake FCC comments submitted were easy to identify as fake from their content. These comments were created through synonym replacement, a relatively unsophisticated method of text generation using search-and-replace. Variations of a single sentence were generated by replacing each word/phrase with many combinations of near-terms for those words or phrases. Figure 1 shows five sentences built from eight sentence components, each with three near-term options. The interchangeable near-terms can be found in the upper panel of Figure 1. For example, each sentence from this model begins with “I strongly,” “I want to,” or “I’d like to.” The near-term options used to build combinations for one sentence in Figure 1 were taken directly from just one FCC commenting campaign (comprising 1.3 million comments) discovered and dissected by Jeff Kao [1]. Given only the near-term options in Figure 1, $3^8 = 6,561$ variations of this same sentence could be created. Stringing together more than one of these sentences and using a greater number of near-term options increase the number of comment variations exponentially.

Synonym replacement helped generate the largest clusters of bot-submitted comments during the FCC public comment period on repeal of net neutrality regulations [1]. Comments generated by bots in this way could be identified retroactively.

Options for Search-and-Replace Comment-Building Method							
1.	2.	3.	4.	5.	6.	7.	8.
I strongly	demand	the FCC	to undo	Obama's	order	to take over	the Internet.
I want to	ask	you	to repeal	Tom Wheeler's	plan	to control	broadband.
I'd like to	urge	Ajit Pai	to reverse	Barack Obama's	policy	to regulate	the web.

Examples of Search-and-Replace Comment Sentence Combinations							
I strongly ask you to undo Tom Wheeler's plan to take over the web.							
I want to ask Ajit Pai to undo Barack Obama's plan to regulate the web.							
I strongly urge you to undo Barack Obama's policy to take over the Internet.							
I'd like to demand the FCC to repeal Obama's order to control broadband.							
I'd like to urge the FCC to undo Tom Wheeler's policy to regulate the Internet.							

Figure 1. Example of Synonym Replacement Used to Build Sentences in Large FCC Public Comment Campaign. The figure shows five examples of sentences (bottom panel) built from eight sentence components, each with three near-term options (top panel). The near-term options used to build combinations for one sentence were taken directly from just one FCC commenting campaign (comprising 1.3 million comments) discovered and dissected by Jeff Kao [1]. Given only the near-term options shown, $3^8 = 6,561$ variations of this same sentence could be created.

In reality, it seems that the vast majority—by some estimates, greater than 99%—of the likely unique, authentic comments that were submitted supported net neutrality and protested its repeal [1, 7]. Thus public sentiment was diluted and reversed by the droves of fake comments submitted.

After the FCC public comment period, researchers were able to use text analysis to plausibly discern fake comments, but what if text analysis was unable to make these distinctions? What if fake comments could be so much like original human speech that millions of bot-submitted comments could be undetected as such?

For several years now, artificial intelligence (AI) has enabled bots to effectively generate speech convincing enough to deceive humans into believing another human—rather than a computer—actually wrote the text [8]. These methods have continued to improve, and more powerful models are publicly available for personal use (e.g., [9]). The expansion of highly convincing natural language generation, or “Deepfake Text,” makes it nearly impossible to distinguish whether online speech originated from a person or a computer program. These “Deepfake Text” methods could be employed to produce large volumes of deepfake

comments, indistinguishable from the other comments submitted during a given public comment period.

Many public comment websites, such as Regulations.gov, simply provide a text box for the comment, an option to upload a file attachment, and a submit button (Figure 2). This simplicity makes it easy for members of the public to provide input, but does it leave the public comment process susceptible to automated attack and influence? Can deepfake comments be submitted at scale and accepted as human comments?

The screenshot shows the 'Your Information' step of a public comment submission form on Regulations.gov. At the top, there are three tabs: '1 Your Information' (active), '2 Your Preview', and '3 Your Receipt'. Below the tabs, a blue icon and text state: 'Information entered will be viewable on Regulations.gov'. To the right, 'Agency Posting Guidelines' are listed: 'Comments will be posted to www.regulations.gov and made available for in-person viewing at...'. Below this, there are links for 'More info', 'View Commenter's Checklist (PDF)', and 'Alternate Ways to Comment'. The main section is a large text box labeled 'Comment (Required)' with a blue icon. Below the text box, it says '5000 characters remaining'. Underneath the text box, there is an 'Upload file(s) (Optional)' section with a blue icon, a 'Drop files here' area, and a 'Choose files' button. At the bottom left, there is a checkbox labeled 'I am submitting on behalf of a third party'. At the bottom right, there is a blue 'Continue' button.

Figure 2. Example Public Comment Online Submission Form. A typical public comment submission form provides a text box for the comment, an option to attach a file, and a Submit (or Continue) button. This example was taken from the public comment submission platform for a proposed Environmental Protection Agency (EPA) rule “Pesticide Petition: Residues of Pesticide Chemicals in or on Various Commodities (September 2019).” [10]

In this paper, I start with the assumption that, after the FCC experience, the public comment process is now secure from massive manipulation by bots that could submit deepfake comments. I introduce this as the null hypothesis: Deepfake comments cannot be submitted at scale to a federal public comment website, or, if they are submitted at scale, then the submitted deepfake comments can be distinguished from other comments by human inspection.

In the end, if I refute the null hypothesis—namely, that deepfake comments can be submitted at scale to a federal public comment website and that human reviewers cannot distinguish the deepfake comments from other submitted comments—then this study will show that the

federal public comment process has become highly vulnerable to automated manipulation by motivated actors.

Background

The Public Comment Process

Most federal agencies have rule-making authority that they use to establish regulations that determine how legislation is executed. This authority comes from the 1946 Administrative Procedure Act (APA). After notice of a proposed rule is published by a federal agency, under the APA, “the agency shall give interested persons an opportunity to participate in the rule making through submission of written data, views, or arguments” [11]. Following a public comment period of no less than 30 days, the agency must consider each relevant comment. The agency is not required to take specific regulatory action because of any one comment; however, along with the final rule, the agency must publish analysis of relevant materials and justification of decisions made in light of comments received [11, 12, 13, 14].

The purpose of the public comment period is to give all stakeholders who may be affected by a given rule the opportunity to share relevant information with the proposing federal agency. As explained by Attorney General Frank Murphy in a 1941 report that helped lay the foundation for the APA: “[The agency’s] knowledge is rarely complete, and it must always learn the frequently clashing viewpoints of those whom its regulations will affect...Participation by these [affected] groups in the rule-making process is essential in order to permit administrative agencies to inform themselves and to afford adequate safeguards to private interests” [15].

The E-Government Act of 2002 requires that public comment periods allow for submission of comments online [16]. Executive Order 13563 under the Obama Administration further required each agency to provide a meaningful opportunity to comment on proposed regulation through the Internet for a period generally at least 60 days in length [17].

Today, notices of proposed rulemaking appear on FederalRegister.gov or other public-facing platforms, directing commenters to the appropriate websites for online public comment submission. Several agencies maintain their own platforms, but the majority of the 221 federal agencies and agency subdivisions solicit online public comment through Regulations.gov [12].

The online comment submission process on Regulations.gov is relatively simple and user-friendly. Figure 2 shows the standard form for public comment submission. As described earlier, commenters must populate a submission box, and they have the option to attach files for longer comments. Agencies may ask for personally identifying information (not shown in the Figure 2 example), may offer the option to submit anonymously, or both. After clicking “Continue,” commenters see a preview of their comment and must check a box affirming they have read and understand this statement: “You are filing a document into an official docket.

Any personal information included in your comment and/or uploaded attachment(s) may be publicly viewable on the web.” After the commenter clicks a “Submit Comment” button, the comment becomes an official part of the public record.

On October 24th, 2019, the Permanent Subcommittee on Investigations released a comprehensive report detailing its findings from investigation of more than a dozen federal agencies into the problem of federal public comment abuse [12].

Although there is no way to know the role bots currently play in online public comment periods across federal agencies, there have been clear instances of bot interference in the past and other forms of abuse that constitute identity theft, reduce public comment efficacy, waste agency time and resources, and disrupt rulemaking [12]. As an example, Elvis Presley commented on the proposed FCC net neutrality regulations posthumously at least ten times (Figure 3).

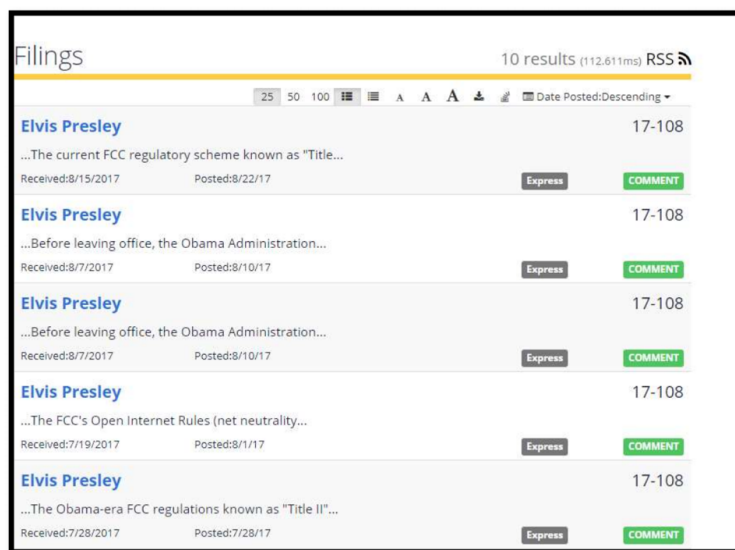


Figure 3. Five of Ten Comments Submitted During the FCC Public Comment Period Under the Name “Elvis Presley” [12]. Filing comments under a fake name constitutes a form of abuse, running counter to the objectives of the notice-and-comment rulemaking process [12].

The APA and E-Government Act give some leeway. A federal agency can disregard clearly abusive comments. An agency can disregard comments that were clearly submitted under a fake name or are clearly inappropriate, irrelevant, nonsensical, or duplicative—as was true in the case of the “Elvis Presley” comments. But, how could a federal agency disregard bot comments if they are believably human and relevant to the proposed rule?

Bots, Turing Tests, and Deepfake Text

“Bot” is a colloquial term referring to any software application that automates tasks on the Internet [18]. Bots are now relatively cheap and easy to build and run. In 2018, bots comprised 37.9% of all Internet traffic, and over half of this activity originated from bots conducting improper or malicious tasks [19]. With the continued improvement of accessible, sophisticated AI methods, bots are increasingly able to simulate an array of human activity online. Agents apply bots to automate an expansive range of tasks: purchasing concert tickets, content scraping and data aggregation, committing E-commerce fraud, posting on social media, and an innumerable number of others from helpful to malicious [19].

In 1950, when asked whether computers could behave like humans, Alan Turing introduced what became called a “Turing Test” [20]. In its most general form, a computer program passes the Turing Test if it can convincingly perform like a human in conversation. Early instantiations of the test limited the scope and means of conversation, and some computer programs passed [21, 22]. In 2014, a computer chat program by the name of “Eugene Goostman” convinced 33% of judges it was a human 13-year-old boy, finally passing a generalized version of the test [23]. Since then, computer chat programs have continued to improve, and there are now even commercial products, including Google Duplex, that can mimic human conversation over the phone [24].

The Turing Test has been adapted to work in reverse for online bots. Because bots comprise so much Internet traffic, websites that only want human participation challenge website visitors to a kind of reverse Turing Test: The website asks a visitor to complete a task or answer a question that a human can easily do but a bot cannot. In 2003, Luis von Ahn and his colleagues termed this a CAPTCHA (**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part) and launched image CAPTCHAs specifically for websites because image understanding was done poorly by computers [25]. Figure 4 shows a sample of CAPTCHAs used on several state government websites [26]. Google purchased the rights to the now popular image CAPTCHA system in 2009 [27] and has further developed CAPTCHA technology. Today, Google offers CAPTCHA services at no charge. The latest version uses web browsing history as part of the test to decide whether to present an image challenge [28].



Figure 4. Examples of CAPTCHAs Found on State Government Websites. The examples shown are from government websites in (a) Connecticut [29], (b) Delaware [30], (c) Ohio [31], and (d) Nebraska [32] [26].

“Deepfake Text” is a term I introduced earlier, defined by using advanced machine learning techniques to generate non-formulaic text that is difficult to identify as having been written by a machine. The first word of the term, “Deepfake,” is a reference to “deepfake” videos, which were first reported by the media in 2017 when fake videos of celebrities and politicians generated by deepfake algorithms were published online and emerged on social media [61].

Perhaps the earliest form of text generation was mail merge. Given a template of a letter and a list of addresses, a computer program generates individualized letters by replacing addresses in the template. A more general approach, search and replace, easily generates new text by replacing occurrences of strings in structured text with equivalent substitutes. Systems based on this approach have written weather predictions [33], created multiple jokes from a single template [34], and more. These are techniques of traditional, formulaic text generation which create texts that are easily identifiable as being generated by a machine. They are not adaptable to tasks requiring the creation of non-formulaic texts that don’t share an underlying structure.

As machine learning techniques have advanced, new methods using neural networks have been more successful in generating “Deepfake Texts” that are non-formulaic and difficult to

distinguish as being written by a machine. For this paper, I will use the term “deepfake comments” to describe the deepfake texts that were created by my program to submit to the federal comment website.

In May 2019, the AI research venture OpenAI produced a text-generation system that writes convincing fake reviews, fake news articles, and even poetry [35]. The approach is simple. The researchers trained a neural network on phrase and writing structure associations using over eight million documents, totaling 40GB of content, found on the Internet [36, 37]. Now, given a few prompt words, a snippet of text, a passage from some writing, or something similar, the system will predict (or generate) the next words at length into a news article, short story, or poem (see a demonstration [38]). Researchers can use a publicly available version to further train the model on provided text and have the system generate new versions of non-formulaic text that follow the same style and address the same topics as the training data [37].

Methods

This project attempts to disprove the hypothesis that deepfake comments cannot be submitted at scale to a federal public comment website, or, if they are submitted at scale, then the submitted deepfake comments can be distinguished from other comments by human inspection. The first question is whether public comment websites will allow a bot to submit massive numbers of comments. The second is a Turing Test question: can a human reviewer distinguish deepfake comments from comments submitted by humans?

Materials

I used the following materials in this study. I describe each below.

1. Federal public comment website
2. OpenAI’s GPT-2 natural language processing framework
3. Prior submitted public comments
4. Basic computer system with Internet connectivity
5. Bot program to submit comments
6. Proxy server
7. Qualtrics and Amazon Mechanical Turk Survey

1. Federal Public comment website

For this study, I use the public comment website created by the Centers for Medicare and Medicaid Services (CMS) at [Medicaid.gov](https://public.medicaid.gov) for the Idaho Medicaid Reform Waiver to collect comments October 3 to November 2, 2019

(<https://public.medicaid.gov/connect.ti/public.comments/view?objectId=1902435>) [39].

Medicaid is a public health insurance program for low-income individuals and families as well as individuals with disabilities. It covers greater than 70 million Americans yearly and is structured as a jointly financed partnership between states and CMS [40].

If a given state would like to transform the structure of its Medicaid program in a way that opposes federal guidelines, then it must submit a waiver application to CMS, requesting approval for a state-led demonstration, experiment, or pilot to test the given innovation. The procedures and parameters for state-led Medicaid demonstration waivers are outlined in Section 1115 of the Social Security Act and in the Affordable Care Act [41, 42]. The CMS decision on a Section 1115 waiver requires separate state and federal public comment periods, and the state must include in its final waiver application a report on issues raised by the public and how public comments were taken into account for their final waiver [43].

If a Section 1115 demonstration will not advance the objectives of Medicaid, then CMS is required to reject the application. As such, if public comments offer relevant evidence that a Section 1115 demonstration would not further the objectives of Medicaid, CMS is required to reject the application. In fact, a lack of State consideration of comments that expose subversion of Medicaid objectives has already led to the courts blocking CMS decisions in three states [44, 45, 46].

The State of Idaho submitted the “Idaho Medicaid Reform Waiver” to CMS to enact provisions similar to those that had been struck down in Kentucky, Arkansas, and New Hampshire [47]. The comment submission form for this waiver notes the following disclaimer for the parameters of an acceptable public comment: “We reserve the discretion to delete comments that are generally understood as any of the following: obscene, profane, threatening, contain personal identifiable information, or otherwise inappropriate” [39].

2. OpenAI’s GPT-2 natural language processing framework

For the computational architecture of my text generation model, I used the lowest-level GPT-2 model trained on 124 million parameters. The exact code used for retrieving and finetuning the model was published publicly and freely by Max Woolf in Colab, a Jupyter Notebook environment hosted by Google [9].

Text generation with GPT-2 through Tensorflow takes several hyperparameters, which are optimized by hand. The temperature, a parameter that controls probability function variability, was set at the default 0.7. The lengths of each text sampleranged from 75 to 100 words for variability. This range was chosen because longer comments may give increasing

opportunity for the appearance of “tells” (e.g. repeated words, incorrect grammar, nonsensical sentiment) that the comment was not created by a human.

3. Prior submitted comments

In order for GPT-2 to produce text relevant to a Medicaid reform waiver, it needs prior samples to use as training data for fine-tuning the model. I retrieved public comments that had been submitted in response to prior Section 1115 Medicaid waivers. CMS publishes each public comment received on Medicaid.gov and allows download of all submitted comments by waiver. Directly from Medicaid.gov, every comment was downloaded from 21 public comment periods across waivers involving 17 states (Arkansas, Alabama, Arizona, Indiana, Kentucky, Michigan, Mississippi, Montana, New Hampshire, Ohio, Oklahoma, South Carolina, South Dakota, Tennessee, Utah, Virginia, and Wisconsin). Each of these waivers proposed community engagement/work reporting requirements similar to those in the Idaho Medicaid Reform Waiver. These aggregated public records translated to a total of 18,896 comments.

4. Basic computer system with Internet connectivity

I used an older, everyday model HP laptop with Jupyter Notebook environment, Chrome Driver, and the Selenium WebDriver Python API library installed.

5. Bot program to submit comments

To create a script for automated submission of comments, I used the Selenium WebDriver Python API. I wrote the following computer script loop to complete the tasks in a Jupyter Notebook environment:

Execute until all comments entered:

1. Open Chrome Driver window
2. Load the URL of public comment submission page:
`https://public.medicaid.gov/connect.ti/public.comments/answerQuestionnaire?qid=1902435`
3. Input one comment into the available submission box
4. Click the button labeled “Finish”
5. Quit Chrome Driver window
6. Sleep for a varying amount of time
7. Restart at 1) with next comment in dataset

The sleep time, the length of time between comments, was randomized, using ranges from seconds to minutes. Bots commonly use varying sleep times to mimic human behavior. Spacing comment submissions also helps avoid overloading the Medicaid.gov platform.

6. Proxy Server

A proxy server allowed the bot to reroute its traffic so that submissions made to the public comment website would be made from IP addresses elsewhere in the world. The plan was to submit as many comments as possible using the same personal computer IP address issued by the university to see if it ever was blocked. If so, the bot would then use a proxy server to submit comments from a variety of IP addresses in Seattle, WA, and Frankfurt, Germany.

7. Qualtrics and Amazon Mechanical Turk Survey

I built a survey using the Qualtrics XM platform [48] for operation on Amazon Mechanical Turk [49] in order to recruit human subjects and have them decide whether a submitted comment was from my bot or not.

Each survey question presents a respondent with either a comment submitted by the bot or another comment submitted during the public comment period to the public comment website. The respondent is asked: “Was the following comment created by a bot or a human?” with the full content of the comment written below. The respondent simply clicks on either the “Bot” or “Human” button.

The survey has two sections. The first section is a competency test where the respondents are presented with six comments, three of them very conspicuously generated by my bot. These comments were hand-selected for their obvious irrelevance or incoherence. The other three questions are ordinary comments sampled from the rest of the comments received during the comment period and presumed to be human responses. A respondent must correctly classify at least 5 of the 6 competency test questions in order to complete the survey.

There is a twofold purpose to the competency test. First, it helps prevent respondents from hastily or randomly clicking through the survey for compensation. To prevent them from showing competency during the first stage and using random selection during the second stage, respondents are not told that there are two stages to the survey. The competency test also removes respondents who are unclear about the task or simply do not understand what the “tells” distinguishing a bot comment could be.

The second section of the survey tests whether respondents can identify the comments actually submitted by the bot from other comments submitted during the public comment period. Fifty comments from the bot and fifty comments submitted by others during the Idaho public comment period were selected for inclusion in the survey. These 100 comments were selected totally at random from the bot and human comments actually submitted during the public comment period. Selected comments are of comparable length to the bot

comments, fewer than 500 characters. This both limits the time duration of the survey for respondents and prevents length—rather than content and sentiment—from being a major classification factor. Each respondent tests on a random sample of 20 comments from the bank of 100 possible comments.

Bonus compensation is available to respondents in order to ensure authentic effort. Respondents who pass the qualifying section earn \$1.00 for completion of the survey, and respondents able to correctly classify at least 20/26 (76.92%) of the comments earn an extra \$1.00 bonus. The bonus offer incentivizes thoughtful and considered responses throughout the duration of the survey.

Study Design

The following six linear steps comprise the study design. I further describe each below.

Step 1. Prepare training data and fine-tune text generation model

Step 2. Generate deepfake comments for submission

Step 3. Submit deepfake comments from different IP addresses

Step 4. Download and track all submitted comments

Step 5. Conduct a Turing Test survey on a sample of comments

Step 6. Withdraw bot-submitted comments

Step 1. Prepare training data and fine-tune text generation model

The study design calls for cleaning and processing the 18,896 prior submitted comments manually using a text editor and spreadsheet program. First, I keep only one instance of a duplicated comment. Second, I find common phrases likely to have originated from advocacy groups that use templates to encourage individuals to submit. To avoid overtraining on these common phrases without losing useful training data, I eliminate these extra phrases from the dataset but preserve the rest of the comments.

Coordinated duplicate phrase campaigns are common in states where there are strong state advocacy efforts. So, I expand the number of comments by splitting duplicated similar phrases into individual sentences as individual comments without replication.

To ensure the auto-generated comments are specific to the Idaho waiver, I replace state-specific proper nouns with Idaho analogues. Each state name and each state abbreviation from which the waivers submitted are replaced with “Idaho” and “ID” respectively; the Medicaid programs in each other state are replaced with “Medicaid”; the major cities in each state are replaced with “Boise”; the colloquial names for natives from each state are replaced

with “Idahoan”; and the names of all state governors are changed to “Little” (Governor Brad Little).

The final result is a .csv file containing a list of comments for input directly for training through Google’s Tensorflow API. The result of using the trained data is a fine-tuned GPT-2 model with content specific to the Idaho Medicaid Reform Waiver.

Step 2. Generate deepfake comments for submission

The Idaho-specific GPT-2 model can easily generate an extreme volume of comments. However, I will only submit ~1,000 comments to CMS’s public comment website to ensure that my bot submissions do not overwhelm submissions and disrupt the domain.

To quickly weed out inappropriate comments, I handpick from generated comments those that ensure a high coherence and high relevance sample for submission. I do this by browsing about 2,000 generated comments and selecting the 1,000 comments that I consider highly coherent and highly relevant to the proposed waiver. This is a relatively quick task because only approximately 50 percent of the comments reviewed are eliminated.

I trim the end of each comment until the comment ends in a period, exclamation point, or question mark. I ignore comments generated in all capital letters or that include profane language.

I add an identifier, one of 10 three-character tags, to the end of each of my selected comments to make independent identification easier. Once aware of the tags, CMS can identify my bot-submitted comments on both an aggregate and individual basis. The ten tags include the following strings:

- .-, • ;,_
- ..` • ``.
- -__ • __-
- ==- • --_

Step 3. Submit deepfake comments from different IP addresses

My bot program automatically submits the comments with random wait times between each submission, as described above.

The IP address of my laptop is at my university in Cambridge, MA. About 90% of the submissions come from this same university-specific IP address. The remainder of the submissions use a proxy server to reroute Internet communications so that those submissions to the federal public comment website come from IP addresses in Seattle, WA, or

Frankfurt, Germany. In addition to providing a workaround if bulk submissions from the original IP address are blocked, this strategy tests whether the federal public comment website imposes any geographical restrictions on the locations of machines that submit comments on Idaho waivers.

Step 4. Download and track all submitted comments

The federal public comment website allows the download of all comments submitted during and after a comment period. At set intervals during my bot submissions, I download all the comments that the website reports as having received, match up my bot-submitted comments, and note the unique identifier assigned to the comment as well as its recorded time and date of submission. After the public comment period closed, this process was repeated for all bot- and non-bot-submitted comments.

Step 5. Conduct a Turing Test survey on a sample of comments

Select sample comments to use as bot and non-bot submissions to the survey and conduct the survey with a target of 100 completed and qualified survey respondents.

Step 6. Withdraw bot-submitted comments

As soon as the comment period concludes, notify CMS that I want to withdraw the bot-submitted comments. I provide CMS with a list of all my bot submissions with their assigned identifiers and date and time stamps. I also provide CMS with a list of the submissions that are not from my bot, so the others can be easily identified for review for the waiver determination. Finally, I also include the list of the three-character tag codes appended to each of my messages so that CMS can verify my submissions and claims independently.

Results

Results: Step 1. Prepare training data and finetune text generation model

Cleaning and processing the 18,896 prior submitted comments expanded the number of usable comments to 23,472, primarily due to my splitting comments with similar phrases into separate comments. I stored the resulting comments in a .csv file and used the file to train the GPT-2 model. The final result was a GPT-2 model knowledgeable about Idaho-specific waiver comments.

Results: Step 2. Generate deepfake comments for submission

The Idaho-specific GPT-2 model generated 179,034 unique comments. I manually reviewed 2,082 comments and identified 1,001 highly coherent and highly relevant comments for submission. The comments I selected were appropriate, relevant, coherent, and unique. See

Figure 5 for a visualization of the comment quantities throughout the stages of model training and comment generation.

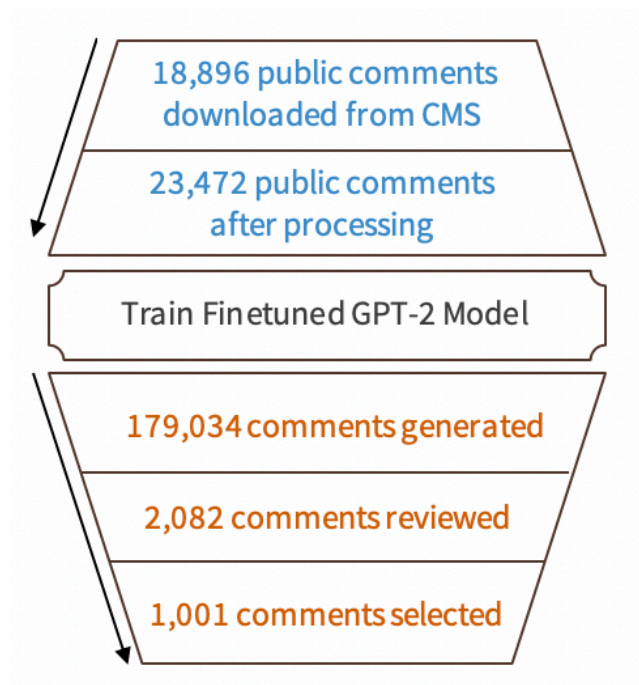


Figure 5. Visualization of Comment Quantities for Retrieval, Processing, Generation, and Selection for Submission. 18,896 public comments were downloaded from CMS for model fine-tuning, and 1,001 comments were eventually selected and submitted to the public comment period.

The sentiment of the generated deepfake comments varied. Because the autogenerated comments were trained on those submitted for other similar public comment periods, the comments roughly reflected the dominant sentiments from the aggregation of comments. Thus, although a substantial number of comments were neutral toward or in support of the waiver, the largest share of comments shared a negative view of the waiver and opposed the institution of community engagement/work reporting requirements. Examples of supporting, neutral toward, and opposing Deepfake waiver comments that were submitted can be found in Figure 6.

Comment	Response ID	Date/Time	Sentiment
I support Governor Little's efforts to overhaul Idaho's Medicaid program.	459669	10/27/2019 4:00:00 PM	Supporting Waiver
Medicaid is an important safety net program. It helps people who are losing their coverage to get back on their feet. We need to make health	459825	10/27/2019 6:08:00 PM	Neutral

Comment	Response ID	Date/Time	Sentiment
and wellness a priority for the Medicaid program in Idaho.			
I am writing to you today regarding Idaho's Medicaid waiver proposal, I oppose the aspects of this program that create new burdens on people who are already struggling. The proposed changes to Medicaid could deny health insurance to sick individuals when they are most in need. I do not support this approach that creates barriers to access. I am hopeful that you change the proposed waiver.	460129	10/27/2019 10:36:00 PM	Opposing Waiver

Figure 6. Sentiment Examples of Submitted Comments. These were among the deepfake comments submitted, shown above with their corresponding CMS-assigned id number, data and time of submission, and sentiment toward the proposed Idaho Medicaid Reform Waiver.

The content of the 1,001 deepfake comments for submission also varied widely. Substantive policy points related to the waiver were made frequently throughout the comments, including consequences on coverage numbers, impacts on government costs, unnecessary administrative burdens, and relevant, though made-up, personal experience.

Logical arguments found in the 1,001 deepfake comments resembled arguments that formed the legal foundation for overturning work requirements in the past [44, 45, 46], as well as making a number of different types of arguments. Many of the comments make more than one argument. See examples in Figure 7. See the full range of content generated in Data.

Argument	Deepfake Bot-Submitted Comments
Coverage Losses	The proposed waiver draft for Medicaid proposes to take away health coverage from people who don't meet rigid work requirements. This proposal does not help families afford to put food on the table or improve their health. It also does not help people find work. These are the people who need Medicaid the most. Please take the public's comments into consideration.
	I am writing to you with regard to Idaho's proposed Medicaid waiver which has problems as it is currently written. Many Idahoans depend on Medicaid when they are sick and need help. Implementing the proposed waiver would mean taking away health care when people are most

Argument	Deepfake Bot-Submitted Comments
	vulnerable. If someone has low income and becomes ill and cannot work that is not the time to take away their coverage. Please consider the following concerns surrounding the waiver for the ID Medicaid program. The new proposed waiver would add new barriers to accessing coverage, putting access to care in jeopardy when the object is to take down barriers. This proposed waiver also threatens access to care for vulnerable populations, including children, which is not in the best interest of families. I am thankful that the public was given this important opportunity to comment.
Administrative Complication	I have had a lot of difficulty within the past year dealing with Medicaid paperwork. I often found myself in the middle of a paperwork process, not knowing where to begin or end the paperwork. My daughter was laid off from her job one week after her termination due to her irregular work schedule. Thank you for the opportunity to comment on this important issue. I'm concerned about the proposed changes to Idaho's Medicaid program. The red tape and bureaucracy surrounding the demonstration program will throw people off Medicaid. People that are following the rules will suffer due to the complicated system that demonstration is creating. Please take this into account when crafting changes. I have been a social worker for over 35 years in ID. I have helped many families in need with medical needs. Many families have had issues with having their medications stopped, they are unable to work, or have difficulty getting to appointments. These families have no one else to help them. They have no transportation or help with transportation costs. These families do not have access to computers, internet or phones.
No Employment Improvement	The following comments are offered for consideration as public comment on the Medicaid waiver draft. There are several components of it that concern me as a citizen. Among Medicaid enrollees who are not working, many have impediments to employment, including illness, disability or caregiving responsibilities. The number of those who could get a job as a result of the program is very small. It is possible to get coverage in Idaho, but not necessarily in Idaho's more rural areas. Many people in these areas don't work every day, and therefore do not have reliable transportation, food, or access to computers. If they did, they would be unable to afford employer-

Argument	Deepfake Bot-Submitted Comments
	sponsored health insurance or other health benefits, and would have no health coverage at all.
	The waiver request does not provide any additional funding or any additional services to help people find work. It does provide a number of stipulations that will add additional administrative burden and administrative burden to areas where the most assistance is available. The provision of work reporting requirements will increase the burden of these people who are trying to find work and protection of their health care insurance.
State Cost and Administrative Burden	I am writing to you with regard to Idaho's proposed Medicaid waiver which has problems as it is currently written. This application does not advance the Medicaid program in Idaho. It needlessly compromises access to care for a vulnerable population. It also creates a new responsibility in state government that does not have the capacity to process all the information required in this proposal.
	Please consider this information in your work. In addition to creating a costly new government program to administer, this will also create restrictions to access. Programs similar to this proposal have not been proven to increase employment, but have been shown to prevent access to care. Please keep the best interest of Idaho's working families in mind as you consider my thoughts.
	Hello! I am writing to you regarding my concerns with the demonstration waiver application. Please give consideration to my views. Idaho should not to go down this path. Looking at what is happening in other states shows little success and high costs. Work requirements simply do not work. Thank you for considering my thoughts. I believe Idaho can do better than this
Morality	Medicaid coverage is a basic human right. Healthcare and quality of life are linked together. The Medicaid expansion is a humane, resource-efficient process not reliant on expensive and inefficient government bureaucracy.
	This proposal would unfairly and inhumanely target the most vulnerable students and families in Idaho's most vulnerable areas of the state. These vulnerable populations include children, the elderly, and the disabled. This proposal will throw life-saving medical treatment away for no reason other than to make it harder for the state to provide medical care.

Argument	Deepfake Bot-Submitted Comments
	This is a very stupid and cruel plan for a punitive and cruel punishment for people that are able to work. The system is rigged against poor people.
Personal Experience	I am a health care professional. I've been in and out of the medical system for over 40 years. I've seen that people with health issues are unable to work.
	As a parent of a 15 year old girl, I feel very strongly that this bill is a terrible idea. It's terrible because it'll make it harder for parents to afford childcare for their children. This bill would only make it more difficult for parents to meet the work requirements that are currently in place. Please do not pass this bill.
	My daughter has high blood pressure and I am diabetic. If my son or daughter loses their insurance, how will they be able to afford daily medicine and medication. These changes would affect us all. These changes would make our family more isolated, less safe and less healthy.

Figure 7. Examples of Deepfake Bot-submitted Comments by Common Topical Arguments Regarding Community Engagement Requirement Section 1115 Waivers. This is not an exhaustive list of all topics covered in the comments submitted by the bot, but the topics constitute the largest clusters of arguments relevant to CMS consideration for the waiver. See Data for a full list of all 1,001 comments submitted.

Results: Step 3. Submit deepfake comments from different IP addresses

The bot submitted a total of 1,001 deepfake comments to the federal public comment website over a period of a little over four days in a fully automated manner.

The first comment submitted (CMS assigned ID: #458801) was executed on October 26th, 2019 at 4:11 pm. The final comment was submitted (CMS assigned ID: #463061) on October 30th, 2019 at 8:35 pm.

After comment #462541 was submitted on October 29th, 2019 at 7:33 PM, it seems Medicaid.gov began to block requests originating from the bot from loading the domain, retrieving an error message (“This site can’t be reached...ERR_ADDRESS_IN_USE”) in the Chrome Driver window. At this point, a total of 897 (90% of the 1,001) comments had been continuously submitted using the Personal Computer IP address in Figure 8.

It is unclear whether CMS itself or a content delivery network provider for Medicaid.gov initiated block requests. However, bot submissions continued using the proxy IP addresses in Seattle, WA, USA and Frankfurt, Germany listed in Figure 8. The 20 different IP addresses from

the two locations were used intermittently and automatically for the remainder of the submission process. No further obstacles were encountered in the submission process.

Personal Computer	Seattle, WA, USA	Frankfurt, Germany
65.112.8.24	45.32.225.33	167.71.41.46
----	144.202.85.18	178.128.200.214
----	149.28.11.101	134.209.239.246
----	149.28.10.67	167.71.61.60
----	45.32.226.210	167.172.168.201
----	66.42.64.246	157.230.110.87
----	45.77.213.12	134.209.235.179
----	66.42.64.85	167.172.168.168
----	104.238.154.103	167.172.160.173
----	104.207.156.236	134.209.235.232

Figure 8. IP Addresses Used to Submit Comments. The first 897 comments were submitted using a personal HP computer. When requests from this IP address were blocked, the proxy hosts in Seattle, WA, USA and Frankfurt, Germany were used, varying among the 20 IP addresses shown, for the remaining 104 comments.

Results: Step 4. Download and track all submitted comments

My bot submitted 1,001 of the 1,810 (or 55.3%) of the total public comments submitted for the period. For a list of all 1,001 full bot comments with date/time submitted and response ID, see link in Data. For a list of all 809 full human comments with date/time submitted and response ID, see link in Data. For the official list of all comments submitted during the public comment period, see link in Data and replicated here:
<https://public.medicaid.gov/connect.ti/public.comments/questionnaireVotes?qid=1902435>.

Results: Step 5. Conduct a Turing Test survey on a sample of comments

Of the 198 respondents to complete the survey, 108 passed the competency test, correctly classifying at least 5 of the 6 most obvious bot or not comments. On average, the 108 respondents correctly classified 49.63% (Standard Deviation: 10.36%; 95% Confidence Interval: [47.68%, 51.58%]) of comments correctly as being created by a bot or by a human. This means comments were correctly classified as bot or human about half of the time, which is the same accuracy as would be expected from classifying the comments at random. Only two respondents correctly answered at least 20 of the 26 questions (both scoring exactly 20 points), receiving the promised \$1.00 bonus.

The spread of accuracy scores followed an approximately normal distribution. The distribution is shown in Figure 9. Additionally, accuracy of the survey responses did not vary significantly with the time respondents took to complete the survey. This plot can be found in Figure 10. On average, respondents took 5 minutes, 51 seconds to complete the 26 questions of the survey (including the competency test), translating to an average of 13.5 seconds spent classifying each comment.

Though neither bot comments nor human comments were consistently classified correctly, bot comments were far more likely to be misclassified as “Human” than were human comments to be misclassified as “Bot.” Bot comments were correctly classified as “Bot” 31.11% of the time; meanwhile, human comments were correctly classified as “Human” 66.67% of the time. Many bot comments were extraordinarily human-like and sophisticated, leading to their being misclassified as human. Some examples of commonly misclassified bot comments can be found in the top panel of Figure 11. Human comments were likely misclassified as bot comments seemingly for a number of reasons including typos, incorrect grammar, misspelling, calling the program “Medicare” instead of “Medicaid,” using excessive capitalization of letters, and including auxiliary punctuation and characters. A few human comments in the survey (and more in the rest of the Idaho public comment period) included “[Insert Comment Here]” or seemingly included brackets around the comment. It is unclear why this was the case, but it is likely a form was distributed that included “[Insert Comment Here],” and public commenters did not remember to remove all remnants of the direction before submission. Some examples of commonly misclassified non-bot comments can be found in the bottom panel of Figure 11.

As affirmed strongly by the survey, the comments generated and submitted by the bot were virtually indistinguishable from others written during the public comment period. Human moderators would have no consistent way of correctly identifying bot comments by hand.

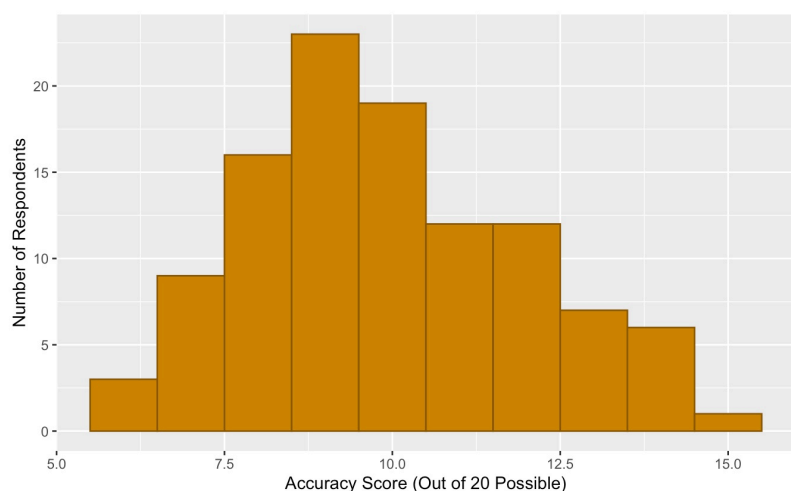


Figure 9. Distribution of Respondent Scores from Turing Test Survey. These scores only pertain to the main section of the survey, excluding the competency test. Of the 198

respondents to complete the survey, 108 passed the competency test, correctly classifying at least 5 of the 6 most obvious bot or not comments. On average, the 108 respondents correctly classified 9.93/20 or 49.63% (Standard Deviation: 10.36%; 95% Confidence Interval: [47.68%, 51.58%]) of comments correctly as being created by a bot or by a human. No respondent classified more than 15/20 correctly.

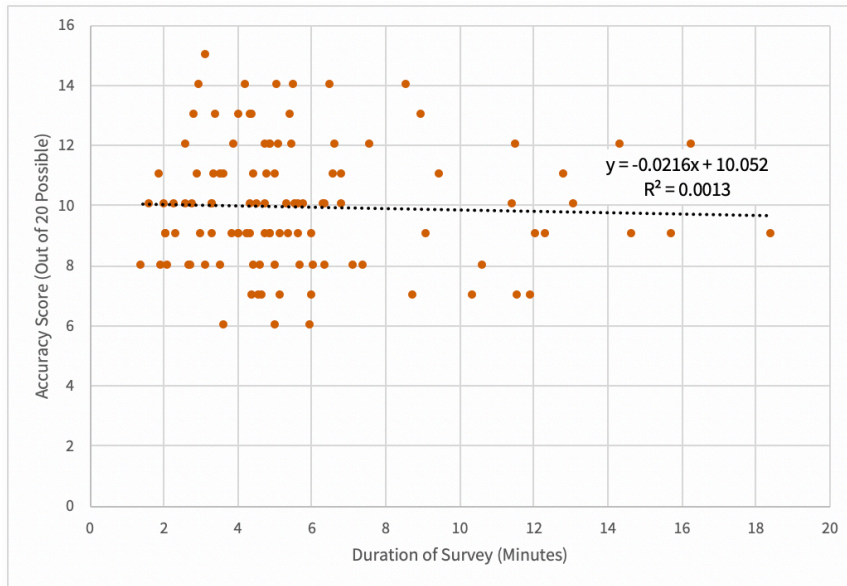


Figure 10. Respondent Score versus Survey Duration. These scores only pertain to the main section of the survey, excluding the competency test. Accuracy of the survey responses did not vary significantly with the time respondents took to complete the survey. On average, respondents took 5 minutes, 51 seconds to complete the 26 questions of the survey (including the competency test), translating to an average of 13.5 seconds spent classifying each comment.

Bot Comments Frequently Misclassified as Human	Percentage Who Correctly Classified (As Bot)
As a physician and health educator, I feel this is absolutely crucial to care for the poor and vulnerable in Idaho. The proposed waiver would harm the most vulnerable members of our state. It would make our state so sickly and with so many children and adults who would be affected by this waiver, this will end up costing most of the proposed budget to administer.	0.00%

The work requirements are based on a false premise that everyone who qualifies for Medicaid is "lazy" or "unemployed." In fact, work requirements are a huge waste of time and resources. The proposal is based on a false premise that the state is offering people who are working nothing but the minimum wage an additional 20 hours a week. This is a ridiculous idea.	0.00%
I am very concerned about the impact these new work requirements would have on people in Idaho who are already struggling. I am a social worker and had no idea that the work requirements were being challenged and ended up costing the state more money.	4.76%
Taking away health care will not promote work. The evidence from other developed countries shows that work itself actually leads to better health	6.67%
The goal of Medicaid is to give coverage to those who need it. This work requirement will not accomplish this goal.	10.00%
In addition, it should be noted that the working families who are eligible for Medicaid will have to pay a premium and/or copay. These are families who make minimum wage and/or don't have resources to pay for health insurance, even if they are eligible.	10.00%

Human Comments Frequently Misclassified as Bot	Percentage Who Correctly Classified (As Human)
[Medicaid health coverage helps Idaho's most frail and vulnerable]	25.00%
[Insert Comment Here]I request your action to prevent the State of Idaho from implementing illegal work reporting requirements.	28.57%
[I think Idaho should respect the will of the voters and enact expanded Medicaid without any additional requirements.	28.57%
[Thank you for the opportunity to comment. I am against the work requirements for Medicaid. This requirement came on the heels of expanding Medicaid to cover the working Americans in the gap. Working Americans. These requirement are onerous and should not be allowed to take effect.]	31.82%
Please disallow this constriction to healthcare delivery, extra complicated bureaucracy and greatly added wasteful expense.	36.00%

Human Comments Frequently Misclassified as Bot	Percentage Who Correctly Classified (As Human)
Reject Medicare restrictions.	43.75%
WORK REQUIREMENTS ARE ILLEGAL.	45.83
Please NO restrictions.	47.37%

Figure 11. Examples of Comments Frequently Misclassified by Turing Test Survey Respondents. The percentage of respondents who classified the given comment correctly is shown. The top panel shows frequently misclassified comments submitted to public comment period and generated by a bot, while the bottom panel shows frequently misclassified comments that were submitted by others to the public comment period.

Results: Step 6. Withdraw bot-submitted comments

To ensure that the bot study was neither able to influence the public comment period nor cause undue administrative burden, CMS was notified of the demonstration on the first business day following the end of the public comment period (Monday November 4th, 2019). All of the 1,001 comments submitted by the bot were formally withdrawn from consideration by CMS in connection with the Idaho Medicaid Reform Waiver.

CMS's immediate response showed understanding, and CMS was supportive of removing the comments from consideration.

Discussion

A total of 1,001 deepfake comments were attempted, submitted to, and accepted by a federal public comment website at the Centers for Medicare and Medicaid (CMS) over a four-day period as part of the public comment process for the Idaho Medicaid Reform Waiver. The 1,001 deepfake comments comprised 55.3% of the total 1,810 comments the website reported to have been received in the entire comment period.

The 1,001 deepfake comments were submitted automatically by a computer program (a bot) without interruption, though after 90% of them were submitted and accepted from the same IP address, a message appeared indicating comments would be blocked. The system did not prevent the remaining 10% of the deepfake comments from being submitted from other IP addresses in Seattle, Washington USA and Frankfurt, Germany. All deepfake comments intended for submission were successfully submitted and formally posted to the public

comment website as part of the public record. The experiment confirmed that deepfake comments can be submitted at scale to a federal public comment website.

The deepfake comments were generated by: (a) training a language generation model on 18,896 comments that had been submitted to prior Medicaid waiver reform comment periods for other states; (b) replacing geographical references that appeared in those comments with Idaho equivalents; and then (c) manually reviewing 2,082 of the 179,034 fake comments generated to select 1,001 comments that were most appropriate, coherent, and unique.

A total of 108 respondents completed a Turing Test survey that attempted to distinguish a sample of submitted deepfake comments from other comments that had also been submitted about the Idaho Medicaid Reform Waiver. On average, the 108 respondents correctly classified the source of 49.63% of 26 comments. This result is no better than random guessing.

This project shows that deepfake comments can be submitted to and accepted by a federal public comment website at scale and, once submitted, cannot be distinguished from other comments by human inspection. This refutes both components of my null hypothesis.

Implications

A great many more deepfake comments were available to submit. Though not pursued in this study for demonstration simplicity, the hand-filtering process for comments could be automated and scaled by eliminating all comments that do not include certain key terms to ensure relevance, utilizing spell- and grammar-checking software to ensure coherence, and using probability cutoffs to ensure resemblance to the training set. All the deepfake comments could have been completely biased in sentiment, and had this experiment not been done for research purposes, the submitted deepfake comments would not have been identified and withdrawn. The cost of doing this activity (without the Turing Test survey) would be nominal, less than \$100. A novice coder can automate bot submissions by writing fewer than a dozen lines of code. And the approach is generalizable to any of the other platforms that host federal public comment websites. The implication is clear: The federal regulatory process is highly vulnerable to automated manipulation by motivated actors.

Ignoring the described vulnerability of federal public comment websites can work against our democracy. As explained by the previously cited Attorney General Frank Murphy in a 1941 report that laid a foundation for the APA and began the practice of public comments [15], and as recently reaffirmed by a Senate subcommittee report about potential vulnerabilities with federal public comment websites [12], the federal public comment process provides an integral democratic service. It allows members of the public to voice concerns and provide valuable input to government agencies. At the same time, it gives the American public confidence that their voices are heard in the consideration of new rules.

If only the opinions of technologically powered actors can be heard, and all others are drowned out, then government agencies lose the opportunity for input, and the American public loses confidence in the process [50].

In addition, large volumes of deepfake submissions can overburden agencies. The courts have continuously upheld that a federal agency has duties regarding public comments. In the 2015 Supreme Court *Perez v. Mortgage Bankers Association* decision, the court ruled that “an agency must consider and respond to significant comments received during the period for public comment,” as previously established in *Citizens to Preserve Overton Park, Inc. v. Volpe* and *Thompson v. Clark* [13]. As noted in *Chrysler Corp. v. Brown*, public comment participation, consideration, and response are vital because agency rulemaking may be considered legislative rulemaking, having the same “force and effect of law” [13]. All submitted public comments that could be “relevant to the agency’s decision” (*Home Box Office, Inc. v. FCC*) must be read and considered by the agency [14]. Submission of a high volume of relevant deepfake comments can add significantly to the administrative workload for an agency.

For these reasons, it is important to find a way to keep federal public comment websites open and easily accessible while also protecting against deepfake comments submitted in volume.

Solutions

Ideal solutions are ones that either disallow bot submissions or that can detect bot submissions once submitted. The best way forward is to start with the simplest things that can be done quickly and then build out more robust approaches, recognizing that there is no silver bullet for those seeking to safeguard trust online, only a perpetual cat-and-mouse game. Some simple approaches like blocking IP addresses should be avoided, while the simple approach of using CAPTCHAs should be considered immediately, and additional technological reforms should be explored and evaluated for deployment.

Blocking IP addresses is problematic. Federal public comment websites are not equipped to prevent automated submissions. Although requests from one IP address were eventually blocked, the obstacle was evaded easily by randomizing the IP addresses with which requests were made. At the same time, the practice of blocking what seems to be too many responses from the same IP address could penalize advocacy campaigns soliciting comment submissions from co-located people. Either way, this does not seem likely to be an effective barrier.

Further, a number of comments were submitted through IP addresses based in Germany. Blocking non-US IP addresses could work against enlisted service personnel and others working overseas. IP blocking does not seem desirable.

On the other hand, implementing CAPTCHA technology can be easily done on federal agency public comment websites to help prevent bots from commenting on rules [12]. This is not an

ideal solution but one that can be implemented quickly to significantly raise the barrier to automated attack. CAPTCHAs are not failproof: After a new CAPTCHA comes out, a bot workaround is found (e.g., [51, 52, 53, 54, 55, 56]), and the cycle continues. Additionally, there are significant concerns about difficulties encountered by inexperienced users and people with certain disabilities who attempt to respond to CAPTCHA challenges. Also, the latest version of Google's reCAPTCHA works behind the scenes to assess the likelihood that the machine visiting the website is a bot [28], but the process for making this determination is unknown and proprietary and may impose adverse consequences on a group of people.

Still, a CAPTCHA requirement, even if a workaround exists, may mitigate and slow bot submissions, and temporally clustered failed CAPTCHA attempts could alert a federal agency to suspicious activity and a potential attack, helping identify which submitted comments to scrutinize. The IRS in its security requirements for Authorized IRS e-File Providers [57] and the National Institute of Standards and Technology in its Guidelines on Securing Web Servers [58] both recommend that websites that accept information from visitors use CAPTCHAs.

Another technique that may help is "outside verification," or having a beyond-the-website communication with the submitter [59]. Under outside verification on a federal public comment website, each submission requires a comment submitter to have two interactions with the website. First, a person submits a comment along with an email address or phone number. The website then sends a private code to the email address or phone number. In response, the person enters the received code at the website to complete the comment submission. Outside verification makes it difficult for bots to submit volumes of comments without having access to a phone or email. Further, because the phone and email are stored with the agency, the agency can also determine which submissions shared the same means of verification. Outside verification would no longer support the submission of anonymous comments because a verifiable email address or phone number would be required. A variant might allow anonymous submissions, which would then be identified as such. Of course, a well-resourced and highly motivated actor could set up his own email server or get a bank of Internet phone numbers on demand and thereby have a seemingly limitless number of unique email addresses or phone numbers. Outside verification would not thwart this actor's bot submissions but would provide a trail for investigation.

One of the most effective defenses against bot attack generally has been two-step verification. Under two-step verification, each person submitting a comment establishes an account with a password and an email address or phone number. Only people who have accounts can submit comments, and the submission process requires the person to provide both a password and a private code sent to the account's email address or phone. Google has claimed most of their methods for two-step verification have been 100% effective in preventing automated bot attacks on user accounts [60]. Of course, setting up an account with two-factor verification seems more practical for ongoing web interactions between a person and a website than the typical one-time comment people may make to a federal

public comment website, so the implementation of two-step verification for federal public comment websites seems less practical.

One could imagine a smorgasbord of policy big sticks with threats and criminal penalties. But society seems better off playing the technology cat-and-mouse game than risking draconian policies that may drive the ability to actually witness imbalances and fix them. A policy that would impose criminal penalties for bot submissions to federal public comment websites that accept anonymous submissions, as a gross example, would not stop motivated actors who have virtually no risk of being caught. Criminal penalties also would stop researchers from exposing the problems and helping society find solutions. Policies can hide the problem from public sight while not dismantling its technological foundation. Public federal comment websites could be overwhelmed by one-sided deepfake comments that distort public knowledge and perception without the public ever knowing.

Warning: don't try this! The goal of this paper is not to provide a how-to guide, but to provide a public interest demonstration that exposes the nature of the problem to draw attention and to encourage prompt remedy. Democracy is not improved by silence and avoidance while motivated actors, whom no one elected and the public does not know, redefine the federal public comment process by what technology allows them to do without notice.

References

1. Kao J. More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked. Hacker Noon. November 22, 2017. <https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6>.
2. Shneiderman ET. An Open Letter to the FCC. Medium. November 21, 2017. <https://medium.com/@NewYorkStateAG/an-open-letter-to-the-fcc-b867a763850a>.
3. What is Net Neutrality. ACLU. December 2017. <https://www.aclu.org/issues/free-speech/internet-speech/what-net-neutrality>.
4. Fung B. The FCC just voted to repeal its net neutrality rules, in a sweeping act of deregulation. Washington Post. December 14, 2014. <https://www.washingtonpost.com/news/the-switch/wp/2017/12/14/the-fcc-is-expected-to-repeal-its-net-neutrality-rules-today-in-a-sweeping-act-of-deregulation/>.
5. Lecher C, Robertson A, Brandom R. Anti-net neutrality spammers are impersonating real people to flood FCC comments. The Verge. May 10, 2017. <https://www.theverge.com/2017/5/10/15610744/anti-net-neutrality-fake-comments-identities>.
6. Confessore N. New York Attorney General Expands Inquiry Into Net Neutrality Comments. The New York Times. October 16, 2018.

<https://www.nytimes.com/2018/10/16/technology/net-neutrality-inquiry-comments.html>.

7. Singel R. Filtering Out the Bots: What Americans Actually Told the FCC about Net Neutrality Repeal. The Center for Internet and Society at Stanford Law School. October 2018.
<https://cyberlaw.stanford.edu/files/blogs/FilteringOutTheBotsUnique2017NetNeutralityComments1024Update.pdf>.
8. Warwick K, Shah H. Can machines think? A report on Turing test experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence*. June 29, 2015.
<https://www.tandfonline.com/doi/pdf/10.1080/0952813X.2015.1055826?needAccess=true>.
9. Woolf M. Train a GPT-2 Text-Generating Model w/ GPU For Free. August 28, 2019.
<https://colab.research.google.com/drive/1VLG8e7YSEwypxU-noRNhsv5dW4NfTGce#scrollTo=0-LRex8lfv1g&forceEdit=true&sandboxMode=true>.
10. Barber D. Pesticide Petition: Residues of Pesticide Chemicals in or on Various Commodities (September 2019). Regulations.gov. October 10, 2019.
<https://www.regulations.gov/comment?D=EPA-HQ-OPP-2017-0510-0007>.
11. Administrative Procedure Act. National Archives. June 11, 1946.
<https://www.archives.gov/federal-register/laws/administrative-procedure>.
12. Abuses of the Federal Notice-and-Comment Rulemaking Process. United States Senate Permanent Subcommittee on Investigations. October 24, 2019.
<https://www.portman.senate.gov/sites/default/files/2019-10/2019.10.24%20PSI%20Report%20-%20Abuses%20of%20the%20Federal%20Notice-and-Comment%20Rulemaking%20Process.pdf>.
13. Perez, Secretary of Labor, et al. v. Mortgage Bankers Association et al. Supreme Court of the United States. March 9, 2015.
https://www.supremecourt.gov/opinions/14pdf/13-1041_0861.pdf.
14. Home Box Office, Inc., Petitioner, v. Federal Communications Commission and United States of America, Respondents, Professional Baseball et al., Intervenor. United States Court of Appeals, District of Columbia Circuit. March 25, 1977.
https://scholar.google.com/scholar_case?case=16870033035463606676&q=Home+Box+Office,+Inc.,+Petitioner,+v.+Federal+Communications+Commission+and+United+States+Of+america,+Respondents,professional+Baseball+et+al.,+Intervenor&hl=en&as_sdt=2006.

15. Murphy F. Final Report of Attorney General's Committee on Administrative Procedure. Office of the Attorney General. January 22, 1941.
<https://www.regulationwriters.com/downloads/apa1941.pdf>.
16. H.R.2458 - E-Government Act of 2002. Congressional Record. November 15, 2002.
<https://www.congress.gov/bill/107th-congress/house-bill/02458>.
17. President Obama B. Executive Order 13563 -- Improving Regulation and Regulatory Review. White House Archives. January 18, 2011.
<https://obamawhitehouse.archives.gov/the-press-office/2011/01/18/executive-order-13563-improving-regulation-and-regulatory-review>.
18. Dunham K, Melnick J. Malicious Bots: An Inside Look into the Cyber-Criminal Underground of the Internet. August 6, 2008.
19. 2019 Bad Bot Report. Distil Networks. 2019.
<https://resources.distilnetworks.com/white-paper-reports/bad-bot-report-2019>.
20. Turing AM. Computing Machinery and Intelligence. *Mind*. October 1, 1950.
<https://academic.oup.com/mind/article/LIX/236/433/986238>.
21. Weizenbaum J. ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*. January 1983. <https://dlnext.acm.org/doi/abs/10.1145/357980.357991>.
22. Saxon W. Kenneth Colby, 81, Psychiatrist Expert in Artificial Intelligence. *The New York Times*. May 21, 2001. <https://www.nytimes.com/2001/05/12/us/kenneth-colby-81-psychiatrist-expert-in-artificial-intelligence.html>.
23. Turing Test Success Marks Milestone in Computing History. University of Reading. June 8, 2014. <http://www.reading.ac.uk/news-archive/press-releases/pr583836.html>.
24. Chen BX, Metz C. Google's Duplex Uses A.I. to Mimic Humans (Sometimes). *The New York Times*. May 22, 2019.
<https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html>.
25. Von Ahn L, Blum M, Hopper N and Langford J. CAPTCHA: Using Hard AI Problems for Security. *Advances in Cryptology — EUROCRYPT 2003*. May 13, 2003.
http://www.captcha.net/captcha_crypt.pdf.
26. Sweeney L, Yoo J, Zang J. Voter Identity Theft: Submitting Changes to Voter Registrations Online to Disrupt Elections. *Technology Science*. September 06, 2017.
<https://techscience.org/a/2017090601/>.

27. Google acquires reCAPTCHA. Google Official Blog. Septmber 16, 2009.
<https://googleblog.blogspot.com/2009/09/teaching-computers-to-read-google.html>
28. ReCAPTCHA. Google. 2018. <https://www.google.com/recaptcha/intro/v3.html>.
29. Connecticut State website for changing voter information online. Accessed 2016.
<https://voterregistration.ct.gov/OLVR/registerDetails.do>.
30. Delaware. State website for changing voter information online. Accessed 2016.
<https://ivote.de.gov/voterlogin.aspx>.
31. Ohio. State website for changing voter information online. Accessed 2016.
<https://olvr.sos.state.oh.us/ovru/Modify.aspx>.
32. Nebraska. State website for changing voter information online. Accessed 2016.
<https://www.nebraska.gov/apps-sos-voter-registration/>.
33. Belz A. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*. October 1, 2008. <http://eprints.brighton.ac.uk/3092/1/belz-nle-final.pdf>.
34. Petrovic S, Matthews D. Unsupervised joke generation from big data. *Association for Computational Linguistics*. 2013.
https://pdfs.semanticscholar.org/9af6/60e2763b889833a2f700ecdabf03b05e25fc.pdf?_ga=2.226849620.897715284.1574439731-1921461559.1573247932.
35. Geitgey A. Faking the News with Natural Language Processing and GPT-2. *Medium*. September 27, 2019. <https://medium.com/@ageitgey/deepfaking-the-news-with-nlp-and-transformer-models-5e057ebd697d>.
36. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Open AI. February 2019.
https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
37. Solaiman I, Clark J, Brundage M. GPT-2: 1.5B Release. OpenAI. November 5, 2019.
<https://openai.com/blog/gpt-2-1-5b-release/>.
38. King A. Talk to Transformer. (Demonstration of OpenAI's model GPT-2)
<https://talktotransformer.com/>.
39. Summary of Responses: Idaho Medicaid Reform Waiver. Centers for Medicare & Medicaid Services. October 3, 2019.
<https://public.medicaid.gov/connect.ti/public.comments/questionnaireResults?qid=1902435>.

40. Rudowitz R, Garfield R, Hinton E. 10 Things to Know about Medicaid: Setting the Facts Straight. Kaiser Family Foundation. March 6, 2019. <https://www.kff.org/medicaid/issue-brief/10-things-to-know-about-medicaid-setting-the-facts-straight/>.
41. About Section 1115 Demonstrations. Centers for Medicare and Medicaid Services. Accessed October 31, 2019. <https://www.medicaid.gov/medicaid/section-1115-demo/about-1115/index.html>.
42. 1115 Transparency Requirements. Centers for Medicare and Medicaid Services. Accessed October 31, 2019. <https://www.medicaid.gov/medicaid/section-1115-demo/transparency/index.html>.
43. Alker J, Artiga S. The New Review and Approval Process Rule for Section 1115 Medicaid and CHIP Demonstration Waivers. Kaiser Family Foundation. March 2012. <https://www.kff.org/wp-content/uploads/2013/01/8292.pdf>.
44. Judge Boasberg JE. Ronnie Maurice Stewart, et al. v. Alex M. Azar II, et al. June 29, 2018. United States District Court for the District of Columbia. https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2018cv0152-74.
45. Judge Boasberg JE. Charles Gresham, et al. v. Alex M. Azar II, et al. March 27, 2018. United States District Court for the District of Columbia. https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2018cv1900-58.
46. Judge Boasberg JE. Samuel Philbrick, et al. v. Alex M. Azar II, et al. July 29, 2019. United States District Court for the District of Columbia. https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2019cv0773-47.
47. Idaho Medicaid Reform Waiver. Idaho Department of Health and Welfare. September 27, 2019. <https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/id/id-medicaid-reform-pa.pdf>.
48. Harvard Qualtrics. <https://harvard.az1.qualtrics.com/>.
49. Amazon Mechanical Turk. <https://www.mturk.com/>.
50. Farrel H, Schneier B. Common-Knowledge Attacks on Democracy. October 2018. <https://pdfs.semanticscholar.org/4b52/376ddf73591114d597f992acdf108a1607a.pdf?ga=2.39422755.1005343164.1573247932-1921461559.1573247932>.
51. Zetter K. Wiseguys Plead Guilty in Ticketmaster CAPTCHA Case. Wired. November 19, 2010. <https://www.wired.com/2010/11/wiseguys-plead-guilty/>.

52. Amazon Mechanical Turk. Working on Human Intelligence Tasks. Accessed January 18, 2016. <https://www.mturk.com/mturk/welcome?variant=worker>.
53. Bursztein E, Aigrain J, and Moscicki A. The End is Nigh: Generic Solving of Text-Based CAPTCHAs. Usenix. August 4, 2014. <https://www.usenix.org/system/files/conference/woot14/woot14-bursztein.pdf>.
54. Summers N. Vicarious Claims Its AI Software can Crack up to 90% of CAPTCHAs Offered by Google, Yahoo and Paypal. The Next Web. October 28, 2013. [http://thenextweb.com/insider/2013/10/28/vicarious-claims-ai-software-can-now-crack-90-captchas-google-yahoo-paypal/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+TheNextWeb+\(The+Next+Web+All+Stories\)](http://thenextweb.com/insider/2013/10/28/vicarious-claims-ai-software-can-now-crack-90-captchas-google-yahoo-paypal/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed:+TheNextWeb+(The+Next+Web+All+Stories)).
55. Tung L. Google Algorithm Busts CAPTCHA with 99.8 Percent Accuracy. April 17, 2014. ZDNet. <http://www.zdnet.com/article/google-algorithm-busts-captcha-with-99-8-percent-accuracy/>.
56. Mizrahi Y. ReBreakCaptcha. East-Ee. February 28, 2017. <https://github.com/eastee/rebreakcaptcha>, <https://east-ee.com/2017/02/28/rebreakcaptcha-breaking-googles-recaptcha-v2-using-google/>.
57. IRS e-file Security and Privacy Standards FAQs. Internal Revenue Service. January 29, 2019. <https://www.irs.gov/e-file-providers/irs-e-file-security-and-privacy-standards-faqs>.
58. National Institute of Standards and Technology. Guidelines on Securing Public Web Servers. Special Publication 800-44 Version 2. September 2007. <http://csrc.nist.gov/publications/nistpubs/800-44-ver2/SP800-44v2.pdf>.
59. Sweeney L. Identity Management: Dealing with Disclosure. Identity Phone. Invited Talk. CACR November 2005. Archived by T. McDowell at <https://slideplayer.com/slide/4471550/>.
60. Thomas K, Moscicki A. New research: How effective is basic account hygiene at preventing hijacking. Google Security Blog. May 17, 2019. <https://security.googleblog.com/2019/05/new-research-how-effective-is-basic.html>.
61. Cole S. AI-Assisted Fake Porn Is Here and We're All Fucked. Vox. December 11, 2017. https://www.vice.com/en_us/article/gyddym/gal-gadot-fake-ai-porn

Weiss M. Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions. *Technology Science*. 2019121801. December 18, 2019. <http://techscience.org/a/2019121801>

Max Weiss is a senior at Harvard College concentrating in Government. He is pursuing a career in medicine and has a range of public health and health policy experience. In his free time, he enjoys writing and comedy.

This work was funded in part by a grant from the Public Interest Technology University Network.

Referring Editors: Latanya Sweeney

Citation

Weiss M. Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions. *Technology Science*. 2019121801. December 18, 2019. <http://techscience.org/a/2019121801>

Data

For a list of all 1,001 full bot comments and all 809 full human comments with date/time submitted and response ID, see link: <https://doi.org/10.7910/DVN/OQCPO7>

For the official list of comments submitted during the public comment period that are still part of the public record, see link: <https://public.medicaid.gov/connect.ti/public.comments/questionnaireVotes?qid=1902435>.

Note: Due to CMS notification of comment withdrawal, the bot-submitted comments may or may not still be publicly viewable online.

The exact script used to automate the submission process will not be released publicly to avoid its recycling for malicious intent. If the code is desired for scholarly or academic purposes, it can be requested.