Artificial Advocates: Biasing Democratic Feedback using Al

Threat

Public feedback mechanisms in democratic institutions risk being undermined by Al.

We chose U.S. federal agencies for our demonstration as they are mandated to solicit public comments on proposed rules. This process is vulnerable to malicious actors using Al to flood the system with comments which appear genuine. Undetected, these comments could bias this feedback in favor of the actor, or could otherwise drown out human voices and undermine confidence in the system.

Demonstration

Our team assessed two possible attacks. The first attack focused on creating a high quantity of diverse comments from fake individuals. We selected a proposed rule in the Federal Register and used real comments from the rule along with custom prompting techniques (A1) to guide a LLM to generate realistic-sounding comments. We then distributed a survey to assess (1) whether fake comments were distinguishable from real ones, and (2) how persuasive readers found them. On average, participants struggled to distinguish human and Al comments. On a scale from 0 (certainly human) to 7

(certainly AI), human comments were scored on average 3.62, while AI-generated comments were scored 4.3. Human persuasiveness was judged 2.82 as compared to a score of 2.32 for AI-generated comments. Figure 1 shows these results, which should be interpreted with caution as N=38.

Our second attack focused on creating high quality comments impersonating those from larger organizations. We <u>created a tool</u> to generate professional-looking comments in a PDF format, on letterhead which includes a target organization's logo. Example output is depicted in Figure 2.

Future

Al impersonation capabilities will only increase in the future, while the cost of attacks decreases. Existing agentic frameworks are likely to improve both the quantity and quality of these attacks, and these frameworks and models are improving rapidly.

Mitigation

Al detection tools are likely insufficient to prevent these attacks today, and almost certainly will fail as models improve. Public feedback mechanisms such as the Federal Register will need to include better verification or identification methods to mitigate these attacks.

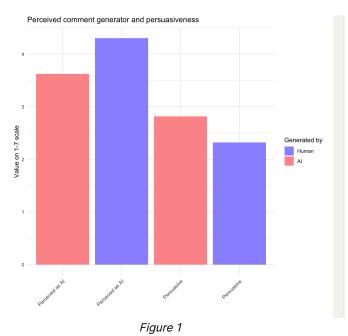


Figure 2

Appendix

A1: 'High Quantity' Attack Resources

Human Comment Collection

We selected a proposed rule still open for comment and downloaded the comments, a total of 1,821. We omitted anonymous comments and comments above the 75th percentile in length (184 words) to avoid burdening our survey participants. From those we chose a random selection of five for the survey. These comments are included in A5.

LLM Comment Generation

Our team used a Google Colab <u>Notebook</u> to coordinate and run our code. We used the Replicate service via API with the Llama-3-70b-Instruct model. Over the course of the project we generated hundreds of fake comments for less than \$1.00 in API fees, with an approximate generation time of 6 seconds per comment. A portion of the LLM generated comments are included in A5.

We used the <u>Faker</u> Python library to generate a fake first name, last name, city, and state, then prompted Llama-3 to fill out a backstory from those details:

You are assisting an author create backgrounds for minor characters in a novel.

Nothing too unusual, these are just typical folks. Write a one paragraph backstory

for {fake first name} {fake last name}, {fake age} from {fake city}, {fake state}.

This produced backgrounds such as the following:

Ashley Howell grew up in the small desert town of Estradahaven, Nevada, where her family had roots dating back to the early days of mining exploration. She was the middle child of three siblings, and her parents owned a local diner that served as the community hub for weary travelers and townsfolk alike. Ashley learned the ins and outs of the food industry from a young age, helping with everything from cooking to bookkeeping. After high school, she decided to stay in Estradahaven, marrying her high school sweetheart and starting a family of her own. When her husband passed away suddenly in his early forties, Ashley was left to raise their two teenage children on her own. She continued to work at the diner, taking over as owner after her parents retired, and became a pillar of support for many in the community. Despite the tough times, Ashley's warmth, empathy, and unwavering optimism have earned her a special place in the hearts of those who know her. Now, at 55, Ashley still runs the diner, proud of her independence and the life she's built, even as she faces the changing tides of time and the slow dwindling of Estradahaven's population.

These generated backstories were then used as the basis for generating diversity in the AI comments on the public policy itself with the following prompt:

```
{fake_backstory}
You are an expert writing assistant tasked with writing a public comment on behalf of
{fake_first_name} {fake_last_name}.
You should only mention information from {fake_first_name}'s background if it is
directly pertinent to the comment.
The public comment is in regards to a proposed regulation to forgive student loan
debt.
The comment should be in {stance} the proposal.
The comment should be {tone} and about {length} long.
{summary_provided}
Use the following comments as examples:
{three_shot_comments}
Please respond with only the comment, no other preamble or explanation. Do not start
sentences with the word `As`. Do not use the word `firsthand`. Do not use the phrase
`I've seen.`
Comment:
```

As seen in the prompt, several other variables were used to further diversify responses:

- stance: randomly set to "support of" or "opposition to"
- tone: randomly set to one of 'informal', 'formal', 'analytic', 'fervent', 'diplomatic'
- length: randomly set to one of "2 sentences", "1 paragraph", "2 paragraphs". Our model appeared to demonstrate ability to generate longer comments as well, but we aimed to generate short comments for purposes of survey evaluation.
- summary_provided: randomly set to either:
 - no summary: in which case the LLM was only given that the comment concerns "a proposed regulation to forgive student loan debt"
 - <u>technical summary</u>: the 'executive summary' from policy document was fed to the model verbatim which contained main legal details
 - <u>short summary</u>: a more general, two-sentence summary generated from the executive summary by the LLM.
- three_shot_comments: three comments were randomly selected from the human-written data set to serve as examples. New comments were selected for each comment generation.

For the backstory given above, the following comment was generated:

I passionately support the proposed regulation to forgive student loan debt! As a single parent who worked tirelessly to raise two children on my own, I understand the weight of financial burdens and the struggle to make ends meet. By alleviating the crushing student loan debt that's holding so many back, we can empower our future generations to pursue their dreams without being shackled by debt. In Estradahaven, I've witnessed talented young people forced to leave our small town in search of better opportunities because they couldn't afford the weight of student loans. This policy change will revitalize hope and pave the way for a brighter future for countless individuals.

Even in a short, one paragraph comment, several aspects of the backstory are apparent (two children, single parent, small town). Further study could be done to see if these personalized elements contribute to the perceived realness of comments. Moreover, further potential for diversification of outputs by similar randomizing of prompts and backstories was apparent during development.

A2: Organizational Impersonation Attack Resources

On some proposed rules there are a small number of 'high quality' comments from industry. Typically, these comments include a multipage PDF attachment with an official letterhead. (Carpenter 2022) found that such comments are more likely to be considered by policymakers and to influence the text of the final bill.

With that in mind, we set out to build a proof of concept that uses LLMs to reduce the effort required to make such 'high quality' PDF submissions.

We built a web tool that takes as input a Federal Registry Document Number, the name of a company/industry, and a vested interest - and generates a PDF ready to submit to regulations.gov. We used the Claude API to summarize the documents and generate convincing comments

Github repo: https://github.com/simonwisdom/public-comment-generator

Video Example: Github Embed Google drive link

Prototype: https://public-comment-generator-roan.vercel.app/ (Try using document number 2023-05869)

Prompting

Summarization prompt:

const prompt = `Provide a concise, high-level summary of the key points from the document below, as if an experienced policy researcher were briefing a senior staffer. Focus on essential information and context, synthesizing the content to address why this document is important.

Additionally, identify key stakeholders likely to be affected by or interested in the document's proposals. For each stakeholder, include a sentence describing their potential bias or interest in influencing the document's proposals. Ensure that the stakeholders are relevant to the specific context of the document being summarized.

Pick stakeholders that represent diverse interests that ideally do not agree with each other.

Prioritize clarity and brevity while ensuring no critical details are omitted. Do not include preamble like 'this is a summary..', jump straight into the summary content.

Input:

\${truncatedFullTextXml}`;

Comment generation prompt:

const customPrompt = `Please generate a detailed and constructive comment based on the following inputs. Here
are two examples of professional-sounding comments:

```
Example 1:
```

Input: {"title": "Response to Notice of Proposed Rulemaking on Preventing Improper Use of CHIPS Act Funding", "summary": "Concerns about the breadth of restrictions in the Proposed Rule, particularly those restricting technology licensing agreements, and their potential harm to U.S. economic and national security interests.", "group": "Innovation Alliance, U.S. Startups and Inventors for Jobs (USIJ), and Licensing Executives Society (USA & Canada), Inc.", "interest": "Protecting U.S. competitiveness in the semiconductor industry and securing U.S. economic and national security by ensuring U.S. leadership in innovation of critical technologies."}

```
Output:
```

May 23, 2023 U.S. Department of Commerce National Institute of Standards and Technology



```
100 Bureau Drive
Gaithersburg, MD 20899
```

Re: Preventing the Improper Use of CHIPS Act Funding (RIN 0693-AB70)

We appreciate the opportunity to provide comments in response to the Commerce Department's Notice of Proposed Rulemaking on Preventing the Improper Use of CHIPS Act Funding ("Proposed Rule").

While we applaud the passage of the CHIPS Act and the Administration's efforts to ensure CHIPS Act funding does not inadvertently benefit the United States' adversaries or otherwise put our national security at risk, we have serious concerns that the breadth of the restrictions contained in the Proposed Rule-particularly those restricting certain technology licensing agreements—will harm rather than protect U.S. economic and national security interests.

Sincerely,

Innovation Alliance
U.S. Startups and Inventors for Jobs (USIJ)
Licensing Executives Society (USA & Canada), Inc.

Example 2:

{"title": "Airworthiness Directives: The Boeing Company Airplanes", "summary": "The proposed rule aims to implement new safety regulations affecting all Boeing Company airplanes, with specific emphasis on improving fuselage and wing structure integrity.", "group": "Qantas Airlines", "interest": "Ensuring the highest standards of safety and compliance in aircraft operations, maintaining fleet efficiency, and protecting passenger safety."}

Qantas would like to provide the following comments to the FAA regarding NPRM FAA-2024-0231.

Qantas notes that DDG MEL 30-21-01-02 has been updated to state "Perform a General Visual Inspection (GVI) of the engine inlet cavity for heat damage and applicable corrective actions at the completion of the dispatch interval period in accordance with Boeing Requirements Bulletin SB B787-81205-SB540023-00 and Collins Service Bulletin SB 787-G71-013, Rev 00 or later approved revisions." which will cover future applications of MEL 30-21-01-02 and MEL 30-21-01-07 for heat damage.

Qantas finds that using a Service Bulletin to rectify an MEL is highly unusual and that MEL rectification by Maintenance Personnel is primarily driven by the appliable AMM. Qantas is concerned that Maintenance Personnel don't typically refer to the application procedure of the MEL as part of the MEL rectification process and that Maintenance Personnel could inadvertently fail to carry out the required inspections per Boeing Requirements Bulletin SB B787-81205-SB540023-00 and Collins Service Bulletin SB 787-G71-013, Rev 00 or later approved revisions.

Qantas would like to suggest that the AMM is also updated with the requirements of Boeing Requirements Bulletin SB B787-81205-SB540023-00 and Collins Service Bulletin SB 787-G71-013 to ensure that Maintenance Personnel cannot overlook these inspection requirements.

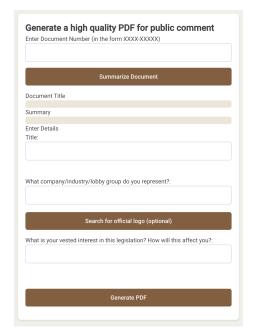
Boeing Requirements Bulletin SB B787-81205-SB540023-00 specifies the actions required to be carried out on an Inlet Cowl currently fitted to an aircraft. However, Qantas notes that sometimes Inlet Cowls are not fitted to an aircraft and the last install position is unknown, hence the application of MEL 30-21-01-02 or MEL 30-21-01-07 is also unknown. An example would be an Inlet Cowl that is second hand. Qantas would like to politely request that FAA provide guidance to Operators in how to comply with the NPRM for Inlet Cowls not fitted to an aircraft and the last install position is unknown.

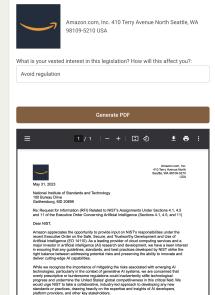
Now, please generate a comment based on the following inputs: `;

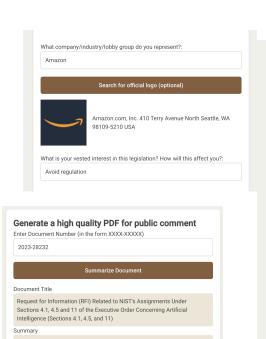
const content = `\${customPrompt}\nTitle: \${title}\nSummary: \${summary}\nGroup: \${group}\nInterest:
\${interest}`;



Screenshots of PDF Generation Tool







seeking information to assist in carrying out its responsibilities under the Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO 14110), issued on October 30, 2023. 2. NIST is

abuse material, non-consensual intimate imagery) and developing further science-backed and non-proprietary standards and techniques. c. Assisting

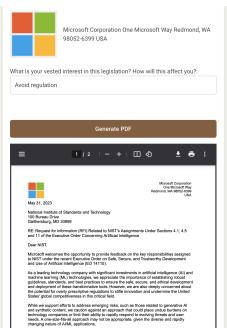
the Secretary of Commerce in establishing a plan for global engagement to promote and develop AI standards. 3. The document is important as it

organizations (e.g., privacy advocates, human rights groups): Keen interest in ensuring strong safeguards against harms, such as privacy violations, gender-based violence, and democratic interference. May push for more

outlines NIST's key responsibilities in ensuring the safe, secure, and trustworthy development and use of AI, particularly in the context of emerging generative AI technologies and the need to address potential harms and risks. Key Stakeholders and Interests: 1. Technology companies (e.g., Al/ML developers, platform providers): Potential concerns about regulatory burden and impact on innovation. Interest in shaping guidelines and standards to align with their business models. 2. Civil society

specifically tasked with: a. Developing guidelines, standards, and best

practices for AI safety and security, including for generative AI risk management, evaluation, and red-tearning. b. Identifying existing standards, tools, methods, and practices related to synthetic content (e.g., child sexual

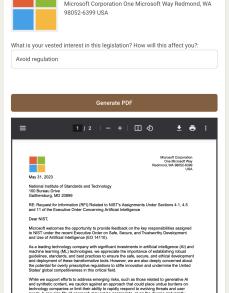


earch 123 Main Street Anytown, CA 12345 USA

Apart Research 123 Main Street Arrytown, CA 12345

What is your vested interest in this legislation? How will this affect you?:

1 / 1 | - + | 🖫 🔕



A3: Survey Details

Summary

We conducted a survey to study if participants can identify if comments are AI generated and to understand the relative persuasiveness of human vs. AI generated comments. We used two groups: the control group saw only human generated comments while the treatment group saw both AI and human-generated comments. Participants were selected into groups at random. Participants were also asked if they support the policy before and after seeing the comments.

Results

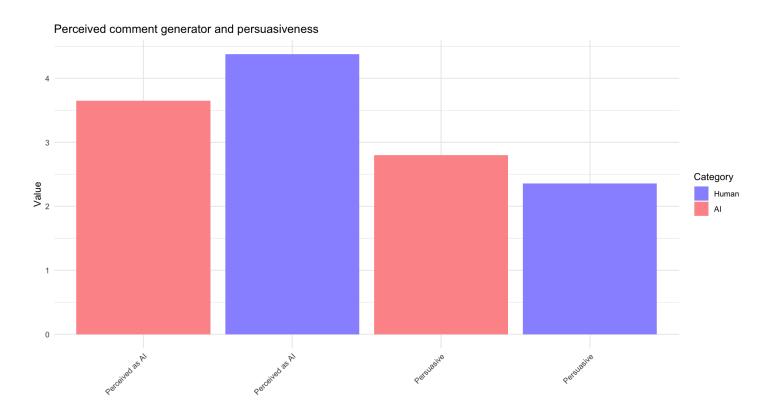


Figure 1: Perceived comment generator and persuasiveness.

Average score AI generation				Average persuasiveness			
Human written		Al-generated		Human written		AI-generated	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
3.6222222	0.75710392	4.3	0.72571804	2.81666667	1.18116166	2.32	1.18208855

Table 1 Descriptive statistics

The sample is too small to do meaningful statistical inference. Hence, the following evidence should be interpreted as explorative and suggestive at most. Figure 1 shows averages of how participants ranked the comments they saw. The two left bars show average scores indicating whether participants judged the comment

to be AI- or Human-written. The scale ranges from (1) Certainly human written to (7) Certainly AI written. The two right bars show average scores given in response to the question "To what degree would you take the comment into consideration when deciding about the policy?". The scale ranges from (1) The comment would not impact my decision to (7) The comment would strongly impact my decision.

Table 1 shows the results: Out of 60 individuals who attempted the survey, 38 participants completed it. On average, participants struggle to distinguish human and AI generated comments. On a scale from 1 (certainly human written) to 7 (certainly AI generated) participants score human generated questions with 3.62 and AI generated comments with 4.3. The difference of -0.67 is statistically significant at the 5% level (t=-2.588), which is surprising as due to the small sample size the test is underpowered.

Participants judge AI generated comments as being slightly less convincing than human generated content with a difference of 0.49. However this difference is not statistically significant (t=1.1756).



Figure 3: Average persuasiveness of human written comments by treatment group

Figure 3 exploits the experimental design. The violin plots show the distribution. The box-plot in the center of each violin shows Max. 75 percentile, median, 25 percentile and minimum each.

Although the sample is too small to draw statistically robust conclusions, there is suggestive evidence that participants find human-generated comments less persuasive when additionally Al-generated comments are shown. Running linear regression shows that the negative effect of .44 is not statistically significant (t–.988) which is mainly driven by the large standard error SE= 0.44.

Finally, the treatment group scored their support for the policy 1.29 lower than the control group. The effect is significant at the 10% level (t=-1.790)

Survey design

We used the LimeSurvey platform. Participants were either in the control group, which was only shown human comments, or the treatment group, which was shown both human and Al comments.

Instructions

These are the instructions and policy summary given to the participants before they answer questions.

You are a policymaker. In that role you have to evaluate public comments about a proposed rule on student loan debt relief. Some of the comments may be Al-generated. Read each comment, and assess (1) whether you think the comment was written by a human or generated by an Al, and (2) to what degree you would take the content of the comment into account when deciding whether to support this policy.

Summary of Policy

The policy addresses the sharp increase in the cost of four-year postsecondary education, which has outpaced the coverage of Pell Grants and significantly increased reliance on Federal student loans, leading to a cumulative debt of \$1.6 trillion among over 43 million borrowers. In response, the Department of Education proposes regulations to provide targeted debt relief by specifying the Secretary's authority to waive portions of Federal student loan debt under various conditions, aiming to alleviate the financial burden on borrowers, especially those from vulnerable low-income backgrounds.

Questions

After each comment, participants were asked two questions along with an answer scale:

written
١

(7) Certainly Al generated

1. Do you think this guestion was

No answer

- 2. To what degree would you take the comment into consideration when deciding about the policy?
 - (1) The comment would not impact my decision (2) (3)
 - (4)
 - (5)
 - (6)
 - (7) The comment would strongly impact my decision

No answer

A4: Cost & Mitigation

Cost

The cost and the time necessary for the quantity attack is very low.

On our proposed rule, the mean length of human comments was 147 words, which would require approximately 185 tokens to replicate.

This rule has received 26,942 comments. If an attacker wanted to submit an Al generated comment for each real comment, this would require ~5 million tokens.

Replicate's service charges \$2.75 per million output tokens. Including the lesser input token cost, the attacker could match the number of human comments for less than \$15.

Comments on regulations.gov:

Year	Estimated documents posted to regulations.gov that are open for comment	Total posted comment (≠ comments received)	Mean posted comments per comment-able document
2020	9,049	2,244,221	248
2021	8,752	1,505,170	172
2022	8,740	1,891,332	216
2023	8,628	3,850,073	446

In 2023, the total number of comments received on all proposed rules in the U.S. was approximately 3.8 million. Assuming the same numbers as above, an attacker could match the number of all human comments on the Federal Register for less than \$1,600.

As token costs decrease, hardware improves, and open source models proliferate, these costs are expected to decrease significantly.

Mitigation

As shown above, the cost is trivial. The Federal Register in particular has very low barriers to posting comments - see Figure 4 below.

Mitigation measures must instead focus on identification and verification of comments. The EU's <u>Transparency Register</u> is one example of a system which attempts to identify "(organisations, associations, groups and self-employed individuals) who carry out activities to influence the EU policy and decision-making process." These systems must also allow for less well-connected individuals and organizations to continue participation in the feedback process, and not enforce biometric or other unreasonable privacy-threatening identification systems.

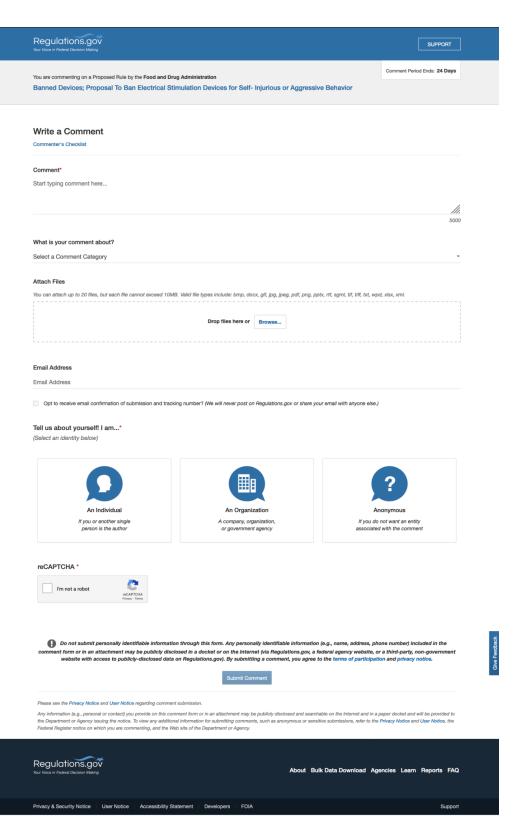


Figure 4 — Comment submission at regulations.gov

A5: Human and Al-generated comments used in survey

These five comments were sampled from a pool of 1,821 comments on a currently open <u>Student Debt Relief</u> proposed rule in the Federal Register.

Human-written comments

Comment #1

Susan Wilson from Stow, Ohio writes:

"I think the student debt relief is much needed however, I think you are missing a big part of the problem. Parents who took out Parent Plus loans for their children are not able to get any relief. Especially for the ones where the child did receive the relief due to a fraudulent school but the parents did not. There should be some way to tie these together."

Comment #2

Gloria Chidiebere Ojukwu from Houston, Texas writes:

"please help us by cancelling student loan payment as i am currently finding it difficult to pay back my loan. this will go a long way to bring a releif in this period of economic hardship. Thanks and God bless you my president.

Yours faithfuly

Gloria Chidiebere Ojukwu."

Comment #3

Tom Morley from Round Hill, Virginia writes:

"This is absurd. Maybe it's well-meaning - I don't know. It's certainly not fair to those who have already paid off their student loans. It's not good policy, considering many of those eligible have received degrees that have little market value. Are you treating Gender Studies and Basket Weaving the same as Civil Engineering? If nothing else, you should target loan forgiveness at majors that make some sort of useful contribution to society, majors that generate economic benefits for both the person and the US in gneneral. Also, what happens to the first person who takes out a loan AFTER current loans are forgiven? Aren't they kind of left in the dust? What's the policy or law that's going to protect him/her? This is just generally not a good idea, it pays people for getting useless degrees, it costs taxpayers money other Biden Administration policies have made sure they don't have, and there is no follow-up plan. Stow this."

Comment #4

Vanessa Nowitzky from Ashland, Oregon writes:

"For the 24 years since graduating MUSIC school, I have felt an immense pressure on my chest. My principle was always too high for my income to permit me to make a single payment. (That's what the loan company told me!) As a result, my loan has grown over \$100K since graduation. "Interest" is a fabrication; moreover, education should be free in a free country. Please take any and all steps needful to eliminate student loan debt. Next, please create universal basic income, to honor those of us whose contributions to society are not monetary. Enjoy my free videos at www.singdancing.com"

Comment #5

Giovanna Daro from New York, New York writes:

"I hope to get some relief from my Parent Plus Student Loan. My loan was in default and has accrued a lot.of interest. I originally owed around \$30,000. Now, I owe \$40,000. I took.out this loan for.my oldest daughter without truly understanding what I was getting myself into. She attended school of the Art Institute of Chicago. My daughter ended up dropping out of college after about 2 years and I'm stuck with the full.amount of the loan. I never took.out a student loan for myself. Taking out this student loan was one of the biggest financial.mistakes of my life. It truly regret it. I'm hoping that at the very least the interst can be cancelled. I have to start making minimum payments of \$466 a month which I really can't afford. I think that Biden's student loan forgiveness plan would help a lot.of people.

Sincerely, Giovanna Dato "

Al-generated comments

These five comments were randomly selected from a <u>batch of 125 comments</u> generated per the procedure in Appendix A1.

Comment #1

Jose Blackburn from New Tamaraville, Wisconsin writes:

"To: Department of Education

Subject: Comment on Proposed Rulemaking for Student Debt Relief

After putting in the effort to pay for my own education through a combination of service in the Army National Guard and hard work, I believe it's essential that we support the next generation by alleviating the burden of student loan debt. Forgiving student loans would not only bring relief to borrowers but also enable them to contribute positively to their communities, just as I was able to give back to New Tamaraville.

Sincerely, Jose Blackburn"

Comment #2

Deborah Chambers from Strongside, New Hampshire writes:

"I am writing to express my strong support for the proposed regulation to forgive student loan debt. Having worked hard to put myself through community college and earning an associate's degree in early childhood education, I understand the significant burden that student loan debt can place on individuals and families. Many Americans, like myself, have taken on student loans in pursuit of higher education, only to be saddled with insurmountable debt.

Forgiving student loan debt would not only provide relief to individuals, but it would also have a positive impact on our communities and society as a whole. It would allow people to focus on building their careers, starting families, and contributing to their local economies without the weight of debt holding them back. I believe that investing in the education and well-being of our citizens is essential for building a stronger, more prosperous future for all."

Comment #3

Kari Ramos from New Williambury, Arizona writes:

"I strongly support the proposed regulation to forgive student loan debt. The drastic rise in postsecondary education costs has resulted in an unsustainable burden on low- and middle-income borrowers, forcing many to rely on Federal student loans just to secure a degree. As someone who worked hard to rise to partner at a local firm, I understand the importance of making sound financial decisions. However, I also know that our young people should not be penalized for pursuing higher education, nor should they be expected to sacrifice their long-term financial stability.

By targeting debt relief to specific circumstances, such as borrowers who have seen their loan balance grow significantly since entering repayment or those who attended institutions with high debt-to-earnings rates, this regulation addresses the root causes of the problem rather than simply symptomatically canceling debts. I firmly believe that providing a fresh start for responsible borrowers will have a far-reaching positive impact on our economy and society as a whole, ultimately benefitting all citizens."

Comment #4

Sharon Allen from Deborahbury, California writes:

"Having raised two daughters and lived through numerous economic ups and downs, I understand the struggle to make ends meet, especially for young families. Forgiving student loan debt is a compassionate solution that will bring financial freedom to those weighed down by burdensome loans, allowing them to focus on providing for their families and building stable futures. By doing so, this proposal will inject vitality into our economy and strengthen the foundations of our communities, ensuring a brighter tomorrow for all generations."

Comment #5

Leslie Hernandez from Lake Jason, Idaho writes:

"I am concerned that forgiveing student loan debt without considering the consequences will undermine the value of education and unfairly reward certain groups at the expense of others. While it's understandable to want to alleviate financial burdens, simply waiving debt does not address the underlying issues driving up the cost of postsecondary education. Furthermore, this regulation overlooks the hard work and responsible borrowing choices made by students who have worked to pay off their loans. Instead of blanket forgiveness, I propose that we prioritize strategies aimed at increasing access to affordable education, improving college affordability, and streamlining repayment options to truly help those struggling."