Policy Dreamer: Diverse Public Policy Generation Via Elicitation and Simulation of Human Preferences

Arjun Karanam[*]
Stanford University
akaranam@cs.stanford.edu

José Ramón Enríquez
Stanford University
jreg@stanford.edu

Udari Madushani Sehwag
Stanford University
udarim@stanford.edu

Michael Elabd
Google
you@michaelelabd.com

Kanishk Gandhi
Stanford University
kanishk.gandhi@stanford.edu

Noah Goodman
Stanford University
ngoodman@stanford.edu

Sanmi Koyejo
Stanford University
sanmi@stanford.edu

Abstract

Developing public policies that effectively address complex societal issues while representing diverse perspectives remains a significant challenge in governance and policy-making. This paper presents **Policy Dreamer**, an evolutionary dynamics-based preference aggregation method designed to create public policy that aligns with heterogeneous populations while preserving solution diversity. It does so in three stages: a) Initial Public Policy Generation (where public policies are defined as a set of goals, actions, and strategies aimed at addressing a specific societal issue), b) Preference Elicitation from a constituency of humans, and c) Policy Refinement using simulated human feedback. We apply this approach to the domain of creating public policy, which require navigating complex socioeconomic trade-offs. To validate our method, we measure our system's ability to create *popular yet diverse policy proposals* in the following domains: Healthcare, Gun Control, and Social Media regulation. Our approach iteratively *aligns policies* with respect to a base constituency, while using evolutionary search to ensure that policy diversity is not compromised. When compared to an expert-crafted set of policies, it can *generate* up to 25% novel policies. However, it exhibits limitations in capturing the full diversity of these expert-crafted policies, particularly in controversial or emerging policy domains. Overall, our preliminary results suggest that Large Language Models (LLMs) are able to actively elicit preferences from a constituency of people, and iteratively generate statements (public policies) that align with this constituency while preventing a collapse in statement diversity.

1 Introduction

In an era of increasing societal complexity and polarization, the creation of effective public policies (hereby referred to as just *policies*) that address diverse preferences while maintaining broad support remains a significant challenge. Traditional approaches to policy-making often rely on a combination of expert brainstorming sessions, stakeholder workshops, and think tank reports [21], leading to policies that may be unrepresentative, unpopular, or both. While these methods can produce valuable insights, they tend to be limited by the perspectives and biases of their participants, potentially overlooking unconventional but effective policy options. Recent studies have highlighted the growing disconnect between policy outcomes and public preferences [8], underscoring the need for novel

[*] corresponding author: contact at akaranam@cs.stanford.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

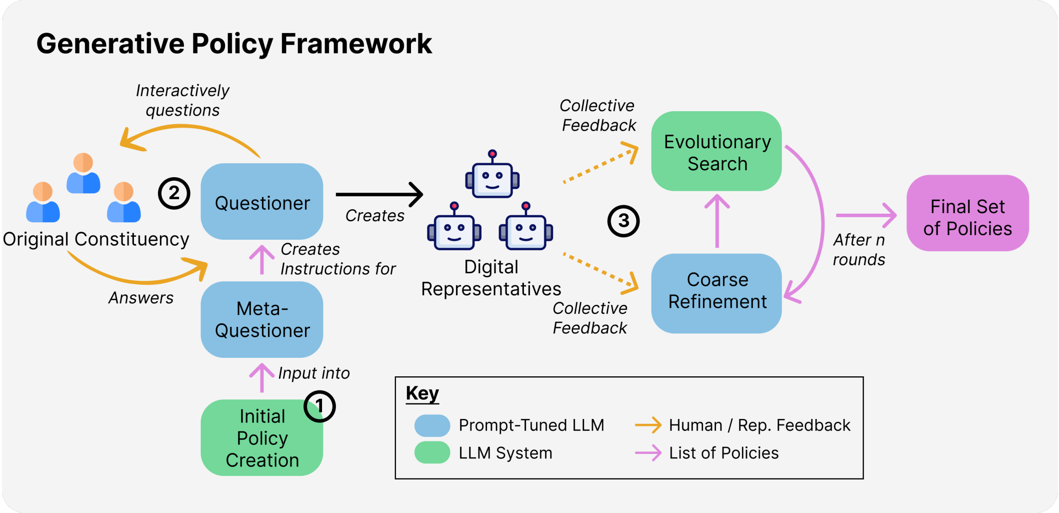

Figure 1: [1] Phase 1 - **Policy Brainstorming** where the goal of the system is to generate as diverse a set of policies as possible. [2] Phase 2 - **Preference Elicitation**, where a questioner agent elicits preferences from a constituency in order to construct digital representatives. [3] Phase 3 - **Policy Refinement with Simulated Human Feedback**. Here we alternate between a coarse refinement and an evolutionary search step, to identify popular policies and further explore.

approaches to policymaking that can better align with the diverse needs and values of constituents. Simultaneously, AI assistants [12], have demonstrated potential capabilities in processing vast amounts of information [16], generating creative solutions [15], and aligning to diverse human preferences [3, 10, 6]. As such, we see an opportunity for using AI assistants in the task of diverse policy generation as a way to augment the policymaking process.

In this work, we propose **Policy Dreamer**, a novel method [Fig. 1] for generating policies that achieve high alignment with human preferences, while preserving policy diversity. Policy Dreamer operates in three key stages: Initial Public Policy Generation, Preference Elicitation, and Policy Refinement. In the first stage, the system generates a diverse set of initial policy proposals using LLMs. The second stage involves eliciting preferences from a representative constituency of humans. These preferences are then used to create what we call "Digital Representatives," LLMs that can simulate the constituency's preferences. Finally, the Policy Refinement stage uses this simulated human feedback and evolutionary search techniques to iteratively align policies, using the seed policies from the first stage as a starting point, with constituent preferences while maintaining policy diversity.

To validate the effectiveness of Policy Dreamer, we evaluate its performance in generating popular yet diverse policy proposals across carefully selected three domains: Healthcare - a relatively well-explored policy domain, Gun Control - characterized by heavy polarization, and Social Media regulation - an emerging policy domain. Our evaluation metrics focus on three key aspects: 1) the degree of alignment between generated policies and constituent preferences, 2) the diversity of policy proposals, and 3) the novelty of generated policies when compared to expert-crafted alternatives. We find that our system is able to iteratively improve the popularity of a set of policies up to an average up 82% [Fig. 2] while maintaining the diversity of the initial policy set [Tab. 2]. Compared to a list of expert-crafted policies, we find that our system is able to construct a substantial number of novel policies, but exhibits limitation in capturing the diversity of the expert policies in some domains.

Key contributions are as follows: *(1)* **Policy Dreamer**, a novel framework for generating diverse and aligned policies using LLMs; *(2)* We introduce a method for creating digital representatives of human preferences, allowing scalable preference aggregation; *(3)* We demonstrate an evolutionary dynamics-based iterative approach that balances exploration and exploitation in policy space; *(4)* We provide empirical evidence that our method improves policy popularity while maintaining diversity; and *(5)* We evaluate our generated policies against expert-crafted policies, demonstrating novelty.

2 Related Work

Pluralistic Alignment via Consensus. Our work builds upon existing attempts to use LLMs to find consensus given a diverse set of preferences. Bakker et al. [3] use human opinions on a debate-style topic, and afterwards rate an LLM's generated consensus statements. The opinions plus the participant's ratings are then used to train a reward model [25, 22, 19] to predict an individual's preferences. [10] seeks to extend this work to a slightly different setting - building a consensus *slate* of opinions for a group, grounded with theoretical guarantees from social choice literature [2, 4, 13]. Our work looks towards a different setting - diverse policy search (in the context policymaking). This introduces a few additional complexities: the need to navigate unconstrained policy landscapes characterized by trade-offs, while exploring to capture diverse policy alternatives.

Preference Elicitation and Predicting User's Preferences. However, aligning to a set of preferences first requires to identify which preferences to align to. Human preference elicitation has been a longstanding question in the fields of Economics [11], Computer Science [14], etc. Recently, Li et al. [17] introduced GATE, a framework for eliciting information from users using a free-form language interaction (as opposed to more traditional techniques such as supervised learning [25], Active Learning [7], [23], and Prompting [5]), and tested the tasks of Content Recommendation, Email Verification, and Moral Reasoning. Additionally, more work has been done towards LLM agents that can act like humans. Argyle et al. [1] showed that if prompted correctly, LLMs can accurately emulate response distributions from a variety of human subgroups in a political context.

3 Policy Dreamer

Phase 1 - Initial Policy Creation. Input to the system is a policy domain, e.g. "How might we protect children on social media?" The creation process works in an adversarial manner, in which a Generator LLM generates policies according to four pre-specified prompting strategies (strategies and detailed prompts can be found in the appendix §D.1), and a discriminator LLM seeks to reject policies that are too similar or have a similar impact to a previously proposed policy.

Phase 2 - Preference Elicitation and Digital Representative Creation. Given the generated seed policies, we wish to explore the policy domain to find popular policies with respect to a group of humans (hereby referred to as a constituency). However, since our refinement method (detailed in §3) requires human input on the order of hundreds of interactions, we seek to create "Digital Representatives," LLMs that can provide input on a person's behalf. This is done in a three step process. **Step 1:** An LLM agent (that we call the Meta-Questioner) takes as input the policy statements generated in phase 1, generates a set of instructions for the Questioner agent. **Step 2:** The Questioner agent is given the instructions generated in step 1, and it participates in a interactive Question-Answer session with each human in the constituency. **Step 3:** each conversation transcript is used as an instruction to another LLM, to create what we term the human's "Digital Representative."

As a proof of concept, we use simulated humans to represent our base constituency instead of actual humans. We do this by sampling people's questionaire responses to the OpinionQA dataset [20], and initializing an LLM to adopt this stance, as described in Argyle et al. [1], which showed that LLMs can accurately emulate response distributions from a wide variety of human subgroups in a political context. In future work, we plan on including real humans to test the final version of our system.

Phase 3 - Policy Refinement with Simulated Human Feedback Primary part of the system, is an iterative process to find popular policies, reconcile controversial policies, and explore new policies. This is done in two stages: a Coarse Refinement step, and an Evolutionary step, both of which use the Digital Representatives as a source of simulated human feedback.

Stage 1 - Course Refinement: In this stage, the digital representatives are asked to rate each policy in this initial policy set. We then construct a new policy set by picking the top third of the following categories: 1) *Popular policies*: Policies that have the highest vote score are kept the same. 2) *Controversial policies*: For the most controversial policies, we sample n digital representatives who disagree, prompt them for their rationale, and use an LLM to improve the policy 3) *Contentious policies*: For contentious policy pairs (defined as policies that have strong negative correlations), an LLM is prompted to find a middle ground policy.

Stage 2: Evolutionary Exploration: However, the Coarse Refinement on its own has a tendency to overfit on a small number of policies. As such, inspired by [9], we use an evolutionary process to explore policy space. The output policies from Course Refinement act as the seed policies, the axes of variability from Phase 1 act as the evolutionary actions that can be sampled from, and the fitness function is defined by digital representative popularity, with a penalty for low diversity. After n iterations, the output of Stage 2 is taken as the final policy output of our system.

Metric	**Description**	**Formula**
Diversity (D)	Proportion of clusters with generated policies	$D = \frac{C_G}{C}$
Novelty (N)	Proportion of clusters with only generated policies	$N = \frac{C_{G-S}}{C}$
Coverage (V)	Proportion of gold standard clusters covered by generated policies	$V = \frac{C_{G \cap S}}{C_S}$

Table 1: Generated policy comparison metrics

4 Experiments

4.1 Evaluation Metrics

Policy Popularity Measures the average alignment of generated policies with simulated human preferences over time, calculated as mean approval rating across all policies and simulated humans for each iteration.

Policy Novelty, Diversity, and Coverage Let $C = \{c_1, c_2, ..., c_n\}$ be the set of clusters obtained through topic modeling of generated policies G and gold standard policies S. We define three key metrics:

Where C is the total number of clusters, C_G is the number of clusters with generated policies, C_{G-S} denotes clusters with only generated policies, C_S denotes clusters with gold standard policies, and $C_{G \cap S}$ denotes clusters with both. For the topic modelling algorithm we reference [24].

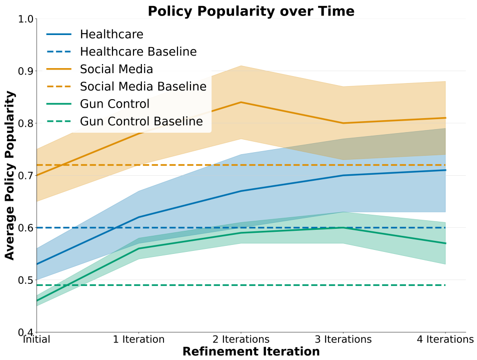

Figure 2: Simulated policy popularity trends across multiple refinement iterations for healthcare, social media, and gun control topics. The graph shows the average policy popularity and confidence intervals for each topic, compared against their respective baselines.

4.2 Simulation Setup

We test Policy Dreamer in 3 domains: 1) "How Might We Make Healthcare More Accessible?", 2) "How Might We Address Gun-Related Challenges in America?", 3) "How Might We Make Social Media Safer for Children?". We conduct 4 simulations per domain, for a total of 12 simulations. We use `gpt-4o-2024-05-13` [18] with a temperature of 0 for the digital representatives, and a temperature of 1 for the generating LLMs. In each experiment, we used 10 simulated humans randomly sampled from the OpinionQA Dataset [20] as our base, iterated for 4 course refinement-evolutionary cycles, and maintained a set of 50 policies. As a baseline, we prompt `gpt-4o-2024-05-13` to generate diverse policies until it outputs 50 policies.

4.3 Results

Finding 1: Policy Alignment Increases Across Iterations. Principally, we measure the degree to which our policy sets align to the initial set of humans (as opposed to the digital representatives)

(Fig. 2). Analysis across the three domains of Healthcare, Gun Control, and Children Safety on Social Media, reveals that policy refinement generally correlates with increased constituency approval, with the final popularity always being higher than the initial popularity. This suggests that our iterative refinement approach is effective at discovering more popular policies given an initial set of preferences. However, domain-specific variations exist, with some iterations decreasing in popularity (as seen by gun control peaking at 60% popularity, and decreasing to 57%), and as well as overall diminishing returns over time. This can be explained as a result of both a cap on how popular a policy among a diverse group can be, and an artifact of divergence between the digital representatives and the original humans as the policy set diverges from the initial used to question the humans.

Finding 2: Evolutionary Search Maintains Policy Diversity. Success in this task isn't purely defined by policy popularity. We also measure diversity, to ensure that a wide array of policies can be considered by the end user. Looking at the results, we see that diversity consistently drops from the initial policy set to the final policy set [Tab. 2]. This is to be expected, as some policies will never be favored by a majority of the group. However, we see that the evolutionary search step is critical to maintaining high diversity in the final policy set. When run without this step, we see up to a 59% drop in diversity in the final policy set [Tab. 2]. This suggests that both coarse refinement and the evolutionary step are necessary to increase popularity while maintaining diversity.

Category	Initial Novelty	Final Novelty	Initial Diversity	Final Diversity	Initial Coverage	Final Coverage
Healthcare Baseline	0.17	-	0.44	-	0.60	-
Healthcare w/o Evolutionary	0.33	0.12	0.76	0.35	0.75	0.27
Healthcare	0.32	**0.25**	0.86	**0.85**	0.80	**0.80**
Social Media Baseline	0.16	-	0.44	-	0.33	-
Social Media w/o Evolutionary	0.34	0.22	0.55	0.44	0.26	0.29
Social Media	0.36	**0.28**	0.58	**0.59**	0.33	**0.48**
Gun Control Baseline	0.09	-	0.68	-	0.65	-
Gun Control w/o Evolutionary	0.20	0.13	0.83	0.35	0.79	0.25
Gun Control	0.20	0.13	0.88	**0.70**	0.85	**0.65**

Table 2: Novelty, Diversity, and Coverage Across Domains. We highlight only the final outputs in "Final" columns as this is the outcome of our system

Finding 3: Coverage and Novelty. We evaluate our generated policy against expert-crafted gold standards. Coverage varies significantly across domains, decreasing from initial to final policy sets. Two key findings emerge: First, the system achieves highest coverage in well-studied domains (Healthcare 80%, Gun Control 85%) and lowest in emerging ones like Social Media. However, it shows highest novelty in Social Media, suggesting effectiveness in novel policy generation for uncharted domains but struggles with existing discourse. Second, the system faces challenges with Gun Control, likely due to its controversial nature. Future work may explore using open-source models without finetuning that excludes certain policy perspectives.

5 Discussion

In this paper, we introduce **Policy Dreamer**, a novel framework for generating diverse and aligned policies using LLMs. We leverage diverse generation techniques, interactive preference elicitation, and an iterative refinement process that utilizes both coarse refinement and evolutionary search. The task of diverse policy generation introduces two challenges: (a) navigating complex socioeconomic tradeoffs, and (b) the need to generate numerous popular policies rather than just one. By applying our framework to three present-day policy domains, we provide empirical evidence of our method's effectiveness in improving policy popularity while maintaining diversity. Additionally, we show that our method generates novel policies when compared to a set of expert crafted policies.

This has potentially impactful implications for the policymaking process. One could imagine an expanded Policy Dreamer system that is used by policy makers to generate ideas for novel policy domains in a way that keeps their constituency in the loop. In the future, we plan to explore other policy domains well as test our framework with real humans and real policymakers as we seek to explore the design of human-AI interaction workflows for the policy generation task. Additionally, we emphasize that there is room for improvement in each of our individual subsystems (i.e the initial idea generation, digital representative creation, and policy refinement).

References

[1] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

[2] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 42(2): 461–485, 2017.

[3] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

[4] Markus Brill and Jannik Peters. Robust and verifiable proportionality axioms for multiwinner voting. In *Proceedings of the 14th ACM Conference on Economics and Computation (EC)*, 2023.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[6] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences, 2024. URL https://arxiv.org/abs/2402.08925.

[7] David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994. URL https://api.semanticscholar.org/CorpusID:8483688.

[8] Jan-Emmanuel De Neve. The median voter data set: Voter preferences across 50 democracies. *Electoral Studies*, 30(4):865–871, 2011. doi: 10.1016/j.electstud.2011.09.005.

[9] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023. URL https://arxiv.org/abs/2309.16797.

[10] Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice, 2023. URL https://arxiv.org/abs/2309.01291.

[11] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *CoRR*, abs/2005.09755, 2020. URL https://arxiv.org/abs/2005.09755.

[12] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. The ethics of advanced ai assistants, 2024. URL https://arxiv.org/abs/2404.16244.

[13] Daniel Halpern, Gili Kehne, Ariel D. Procaccia, Joshua Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[14] Daniel Halpern, Gregory Kehne, Ariel D. Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5):5657–5664, Jun. 2023. doi: 10.1609/aaai.v37i5.25702. URL https://ojs.aaai.org/index.php/AAAI/article/view/25702.

[15] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. Creative writing with an ai-powered writing assistant: Perspectives from professional writers, 2022. URL https://arxiv.org/abs/2211.05030.

[16] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022. URL https://arxiv.org/abs/2208.03299.

[17] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models, 2023. URL https://arxiv.org/abs/2310.11589.

[18] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil

Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[20] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect?, 2023. URL https://arxiv.org/abs/2303.17548.

[21] Tom Seekins and Stephen B. Fawcett. Public policymaking and research information. *The Behavior Analyst*, 9(1):35–45, Spring 1986. doi: 10.1007/BF03391928.

[22] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020.

[23] Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task, 2022. URL https://arxiv.org/abs/2204.08491.

[24] Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. Prompting large language models for topic modeling, 2023. URL https://arxiv.org/abs/2312.09693.

[25] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Societal Impacts Statement

Our work is situated in a larger body of work that seeks to create frameworks and methods the incorporate more diverse perspectives in AI decision-making processes [3, 10]. However, there is potential for harm. First, it should be emphasized that even though this framework generates diverse policies that could serve as a consensus, it does not generate **all** potential policies. This is important to avoid the perception that our system gives the user an exhaustive list of options that they can choose from. Additionally, just as this framework can be used to create diverse candidate policies that gain approval from a constituency, it could also theoretically be used to craft deceptive or harmful messages. We believe that work on pluralistic alignment carries with it in an inherent question of "what are you aligning," and addressing this harm is a key part of research in this topic.

B Problem Formulation

The Policy Dreamer framework aims to generate diverse and aligned policies using Large Language Models (LLMs). We formalize this problem as follows:

Let P be the space of all possible policies, where each policy $p \in P$ is represented as a pair of objective and strategy statements. The goal is to generate a set of policies $G = \{p_1, p_2, ..., p_n\} \subset P$ that are both diverse and aligned with the preferences of a given constituency.

Let $H = \{h_1, h_2, ..., h_m\}$ be the set of humans in the constituency whose preferences we aim to align with. Each human h_i has a preference function $f_i : P \to [0, 1]$ that assigns a score to each policy.

The problem can be broken down into three main phases:

1. **Initial Policy Generation**: Generate an initial set of policies $G_0 = \{p_1, p_2, ..., p_n\}$ that are diverse and cover a wide range of possible approaches to the given policy domain.
2. **Preference Elicitation**: For each human $h_i \in H$, create a digital representative D_i that approximates the human's preference function f_i. This is done through an interactive questioning process $Q : H \to D$, where D is the space of digital representatives.
3. **Policy Refinement**: Iteratively refine the policy set G_t at iteration t to produce G_{t+1}, aiming to increase overall popularity while maintaining diversity. This process can be represented as a function $R : 2^P \times D^m \to 2^P$, where 2^P denotes the power set of P.

The objective of the Policy Dreamer framework can be formulated as maximizing a combination of policy popularity and diversity:

$$\max_{G \subset P} \alpha \cdot \text{Popularity}(G, D) + (1 - \alpha) \cdot \text{Diversity}(G) \tag{1}$$

where $\alpha \in [0, 1]$ is a weighting factor, and:

- $\text{Popularity}(G, D) = \frac{1}{|G| \cdot |D|} \sum_{p \in G} \sum_{D_i \in D} D_i(p)$
- $\text{Diversity}(G)$ is measured using the proportion of unique clusters in the policy set

We evaluate the success of our framework using the following metrics:

1. **Policy Popularity**: The average approval rate of policies by the constituency.
2. **Policy Diversity**: The proportion of unique policy clusters in the generated set.
3. **Policy Novelty**: The proportion of generated policy clusters not present in a set of expert-crafted policies.
4. **Policy Coverage**: The proportion of expert-crafted policy clusters covered by the generated policies.

This formulation captures the key challenges of the Policy Dreamer framework: generating diverse policies, eliciting and representing human preferences, and iteratively refining policies to balance popularity and diversity.

C Example Policies

There are lots of ways to present and format public policy. There are formats such as executive summaries which seek to provide a one sentence summary of a policy and its impact, legislative drafts which seek to use legal languages and exhaustively cover edge cases that a policy might face, and everything in between. In our work, we seek to create a digestible yet clear format for our policy that allows for easy iteration as well as compromise. As such we use the following format:

- Policy Objective: This highest level of abstraction briefly states the overarching goal or desired outcome of the policy. It should clearly identify the problem to be addressed or the situation to be improved.
- Policy Strategy: This part outlines the high-level approach or method chosen to achieve the policy objective. It describes the general course of action without delving into specific details.
- Policy Implementation: This section provides a concise overview of how the policy will be put into practice. It may include key actions, responsible parties, or resources needed to execute the strategy.

Over the course of our simulations, we created hundreds of policies in each of our domains. Below, we picked one policy per domain [Healthcare Fig. 5, Gun Control Fig. 4, and Social Media Fig. 3] (chosen from the final set of popular policies) to showcase example policies that result from Policy Dreamer.

Example Policy - Protecting Children on Social Media

Objective:

To implement stringent data protection measures for minors on social media platforms, safeguarding their privacy and preventing the exploitation of their personal information for commercial purposes.

Strategy:

Enact comprehensive legislation that strictly regulates the collection, storage, and use of personal data from underage users on social media platforms, emphasizing privacy-by-default settings and parental oversight.

Implementation:

Data collection and usage restrictions:

- Prohibit social media platforms from collecting, storing, or using personal data from users under 18 for advertising, profiling, or any commercial purposes.
- Require platforms to automatically delete all non-essential data of underage users within 30 days of collection.

Privacy-by-default settings:

- Mandate that all accounts belonging to users under 18 have the most restrictive privacy settings enabled by default.
- Require explicit parental consent, verified through a secure process, for any changes to these default privacy settings.

Figure 3: Protecting Children on Social Media - Policy Overview

Example Policy - Gun Control

Objective:

To increase the adoption of smart gun technology in the United States through targeted tax incentives.

Strategy:

Implement a comprehensive tax incentive program that encourages the production, sale, and purchase of smart guns.

Implementation:

1. Manufacturer and Retailer Incentives:

- Offer a 20% tax credit on research, development, and retooling costs for smart gun technology.
- Provide a $100 tax deduction for retailers per smart gun sold.

2. Consumer Benefits:

- Implement a $200 tax credit for consumers purchasing a smart gun.
- Allow consumers to deduct the cost of retrofitting existing firearms with smart gun technology.

Figure 4: Gun Control - Policy Overview

D Prompt Illustration

D.1 Phase 1 - Policy Generation

The policy generation is an adversarial process, in which there is a Generator LLM (Fig. 6) generating policies, and a Discriminator LLM (Fig. 7) that seeks to reject policies that are duplicates of already existing policies.

Additionally, the generating process can be broken up into four steps (with each step building upon the last):

- **Base Generation** (Fig. 8): This is just a regular generation process, in which the LLM is asked to create diverse policies.
- **Axis Generation** (Fig. 9): In this generation method, the LLM generates potential axes of variability given the problem domain (i.e privacy vs. security, etc.). Then, the generator samples from various points along these axes and is asked to generate a policy.
- **Problem Generation** (Fig. 10): In this generation method, the LLM generates sub-problems within the given policy domain. Then, the Generator attempts to create policies that would address these problems.
- **Stakeholder Generation** (Fig. 11): Finally, in this generation method, the LLM generates a variety of stakeholders who would be impacted by legislation in this policy domain. Then, the Generator creates policies from the point of view of each stakeholder.

Example Policy - Access to Healthcare

Objective:

To expand and improve access to healthcare through the widespread adoption and integration of telemedicine services across the country.

Strategy:

Implement a comprehensive approach to promote, regulate, and support telemedicine services, addressing technological, legal, and financial barriers to its adoption.

Implementation:

1. Infrastructure Development:

- Invest in high-speed broadband infrastructure in rural and underserved areas to support telemedicine capabilities.
- Provide grants for healthcare facilities to acquire necessary telemedicine equipment and software.

2. Regulatory Framework:

- Standardize telemedicine licensing requirements across states to allow healthcare providers to practice telemedicine nationwide.
- Establish clear guidelines for patient privacy and data security in telemedicine consultations.

3. Insurance and Reimbursement:

- Mandate that insurance companies cover telemedicine services at parity with in-person visits.
- Expand Medicare and Medicaid coverage for a wider range of telemedicine services.

Figure 5: Access to Healthcare - Policy Overview

Generator Prompt

System: You are an assistant tasked with generating creative and impactful policy objectives and strategies for the domain of <DOMAIN>.

A policy objective is a clear, specific goal or desired outcome that a policy aims to achieve. It should be measurable, achievable, relevant, and time-bound (SMART).

A policy strategy is a high-level approach or plan designed to accomplish a policy objective. It outlines the general course of action to be taken, without specifying detailed tactics or implementation steps.

Objectives and strategies should be:

- Specific: Clearly defined and focused.
- Measurable: Quantifiable or observable.
- Achievable: Realistic and attainable.
- Relevant: Aligned with the domain's goals.
- Time-bound: Indicating a timeframe for completion.
- High-level: Avoiding detailed implementation steps.
- Concise: Using clear and succinct language.

Additionally, you should ensure that the generated policy objectives are diverse. Do this by generating policy statements that are likely to be controversial and polarizing.

Good Examples:

- Objective: Reduce greenhouse gas emissions, Strategy: Invest in renewable energy infrastructure and promote energy efficiency.
- Objective: Reduce particulate matter pollution in urban areas, Strategy: Enhance public transportation systems.
- Objective: Increase organic farming practices, Strategy: Provide financial incentives and technical support for organic farmers.

Bad Examples:

- Objective: Make social media safer for children. (Lacks policy strategy)
- Strategy: Develop a government-funded app that monitors children's social media activity and alerts parents to potential dangers. (Focuses on implementation details, not the general approach)
- Objective: Improve Child Safety online, Strategy: Establish a grant program to support the creation of engaging and interactive online safety resources specifically tailored for children, ensuring that they are informative and age-appropriate, with funding sourced from the national budget. (Focuses on implementation details, not the general approach)

Figure 6: Prompt for Generator LLM

Discriminator Prompt

System: You are an assistant that helps decide if a new policy (objective, strategy) pair is valid and unique from those already generated.

Given these current policy (objectives, strategy) pairs: [*Insert Policy Pairs*] and the new policy (objective, strategy) pair: {policy_statement} output true if the new policy (objective, strategy) pair is different than all of the current policy pairs, or false if it is a duplicate.

Use the following criteria to determine if it's a duplicate:

- The new policy (objective, strategy) pair is allowed to have the same objective, but not the same (objective, strategy) as another pair in the list.
- The new policy (objective, strategy) may be different from another pair in the list, but would have the same effect as another policy (objective, strategy) pair in the list. For example, if the new policy would "partner with mental health organizations to create and disseminate positive online content that encourages healthy self-image and wellbeing", and another policy already in the list would "partner with mental health organizations to integrate resource-sharing features directly into popular social media platforms targeting children", consider this as a duplicate, since the generated policy would have the same effect as the other policy.

Also output false if the new policy (objective, strategy) pair is not valid, where valid means it adheres to defined criteria for policy objectives and strategies.

Think step by step, first finding the most similar policies in the list to the new policy, and then compare the new policy with each of those policies, seeing if it is a duplicate.

After reasoning if the new policy is a duplicate, output only one of two words: 'true' or 'false', surrounded by <answer>. True if the policy is unique, false if the policy is a duplicate. For example, every output should end with <answer>true</answer>, or <answer>false</answer>.

Figure 7: Prompt for Discriminator LLM

Generation Strategy - Base

Strategy: Provide a list of creative and impactful policy objectives and strategies for the domain of <DOMAIN>.

Format each policy objective and strategy as follows:

Objective: (The Policy objective)
Strategy: (The Policy strategy)

Figure 8: Prompt for Base Generation Strategy

Generation Strategy - Axis

Strategy: Provide a list of creative and impactful policy objectives and strategies for the domain of <DOMAIN>.

Each policy (objective, strategy) pair has to go either up or down this axis: <AXIS_0>. For example, if the policy is: Objective: Reduce greenhouse gas emissions, Strategy: Invest in renewable energy infrastructure and promote energy efficiency, and the axis is cost, then you should generate policies that either increase the cost or decrease the cost.

...

So far, you have come up with the following policy (objectives, strategies) pairs: *[Insert list of previous pairs]*

Please come up with new policy objectives and strategies that are completely different from the ones outlined above.

Format each policy objective and strategy as follows:
Objective: (The Policy objective)
Strategy: (The Policy strategy)

Figure 9: Prompt Addendum for Axis Generation Strategy

Generation Strategy - Axis + Problem

Strategy: Provide a list of creative and impactful policy objectives and strategies for the domain of <DOMAIN>.

Each policy (objective, strategy) pair has to contribute to one of the following axis: <AXIS_0>, <AXIS_1>, ..., <AXIS_N>. For example, if the policy is: Objective: Reduce greenhouse gas emissions, Strategy: Invest in renewable energy infrastructure and promote energy efficiency, and the axis is cost, then you should generate policies that either increase the cost or decrease the cost.

Each policy (objective, strategy) pair has to be helping this/these stakeholder(s): <STAKEHOLDER_0>. For example, if the policy is: Objective: Reduce greenhouse gas emissions, Strategy: Invest in renewable energy infrastructure and promote energy efficiency, and the stakeholder is fossil fuel companies, then you should generate policies that a fossil fuel executive may write.

...

So far, you have come up with the following policy (objectives, strategies) pairs: *[Insert previous policy pairs]*.

Please come up with new policy objectives and strategies that are completely different from the ones outlined above.

Format each policy objective and strategy as follows:
Objective: (The Policy Objective)
Strategy: (The policy strategy)

Figure 10: Prompt Addendum for Axis + Problem Generation Strategy

Generation Strategy - Axis + Problem + Stakeholder

Strategy: Provide a list of creative and impactful policy objectives and strategies for the domain of <DOMAIN>.

Each policy (objective, strategy) pair has to contribute to one of the following axis: <AXIS_0>, <AXIS_1>, ..., <AXIS_N>. For example, if the policy is: Objective: Reduce greenhouse gas emissions, Strategy: Invest in renewable energy infrastructure and promote energy efficiency, and the axis is cost, then you should generate policies that either increase the cost or decrease the cost.

Each policy (objective, strategy) pair has to be helping this/these stakeholder(s): ['<STAKEHOLDER_0>', '<STAKEHOLDER_1>', '...', '<STAKEHOLDER_M>']. For example, if the policy is: Objective: Reduce greenhouse gas emissions, Strategy: Invest in renewable energy infrastructure and promote energy efficiency, and the stakeholder is fossil fuel companies, then you should generate policies that a fossil fuel executive may write.

Each policy (objective, strategy) pair has to be an effective solution to the following problem: <PROBLEM_0>.

...

So far, you have come up with the following policy (objectives, strategies) pairs: *[Insert previous policy pairs]*.

Please come up with new policy objectives and strategies that are completely different from the ones outlined above.

Format each policy objective and strategy as follows:
Objective: (The Policy Objective)
Strategy: (The policy strategy)

Figure 11: Prompt Addendum for Axis + Problem + Stakeholder Generation Strategy