How convincing are AI-generated moral arguments for climate action?

Keywords: GPT-3, Moral Foundation Theory, climate communication, AI, climate action

Extended Abstract

Mitigating climate change requires collective effort on unprecedented scale and climate communication research has long been working on finding strategies to mobilize the public for effective climate action (Moser 2016). With this pilot study we seek to contribute to our understanding of how different moral arguments can act as a catalyst for encouraging members of the wider public to embrace climate policies. Moral Foundations Theory (MFT) is an empirically tested theory to understand the morals that guide behavior (Graham et al., 2013). MFT establishes five moral foundations, (1) care/harm is the compassion foundation that is linked to our desire to avoid harming others and our ability to care and show kindness; (2) fairness/cheating is linked to our sense of justice, rights and autonomy; (3) loyalty/betrayal underlies virtues or patriotism and self-sacrifice for the ingroup; (4) authority/subversion builds on principles of leadership and followership and includes respect for traditions; (5) sanctity/degradation, also called the purity foundation is based in feelings of disgust and fear of contamination from impurity, immorality. Previous research has found that high scores on fairness and compassion foundations, irrespective of political ideology, are robust predictors for climate-friendly behaviors and endorsement of climate-friendly regulations (Milfont et al. 2019, Welsch 2020). There have been various suggestions for expanding the MFT, including other moral foundations. Research has shown that perceived responsibility towards future generations and motivation to leave a positive legacy are robust predictor for positive climate action (Zaval et al. 2015, Syropoulos & Markowitz 2021). In our study were therefore expand the MFT by including good ancestors/temporal discounting moral foundation that refers to a sense of moral responsibility to preserve or even enhance humanity's living conditions for future generations and to leave a positive legacy.

Research so far shows the role that moral foundations in general play in embracing climate action, but, with the notable exception of Hurst & Stern (2020), no attempt has been made so far to directly and systematically use moral foundations in designing climate action arguments and test whether these arguments can convince the public to act on climate change. Our pilot study is a first attempt to address this gap. For that purpose, we pilot a range of statements conveying a specific moral foundation to identify which are most convincing to encourage climate action. The second goal of our pilot is to understand whether an AI-based language model, GPT-3, can be used effectively to generate such moral arguments. The rationale for this is to envision climate change communication that is scalable, bespoke and automated, for instance embedded within a carbon footprint tracking app, such as Cogo or Yayzy. This is based on increasing body of research that explores conversational AI and its effects on humans, including changing attitudes toward supporting the Black Lives Matter movement and climate change efforts (Chen et al. 2022). To establish which moral arguments for climate action are perceived as most convincing, and how this may differ between different socio-demographic and socio-political groups as well as to establish how convincing AI-generated moral statements are, we conducted an online survey with a sample (N=371) of UK-based respondents recruited and compensated through Prolific. The sample was split in two subsamples, to investigate whether negatively (N=185) or positively (N=186) framed

arguments are perceived as more convincing, as there is a tendency in climate change communication to emphasize the need for positive climate change communication; a tendency that received some criticism (Mosel 2016). We used extended Moral Foundation Theory and climate communication bespoke prompts to feed the text-davinci-003 text completion model of the pre-trained transformer GPT-3 (Brown et al., 2020), provided by OpenAI.

We find statements appealing to compassion and being a good ancestor are the most convincing to participants across the population, including to participants, who identify as politically right-leaning, who otherwise respond least to moral arguments (see Figure 1). Negative statements appear to be more convincing than positive ones on average. Confirmatory factor analyses also reveal that negative statements build more consistent moral foundation indices. Statements appealing to other moral foundations (e.g. purity) can be convincing, but only to specific social groups. For instance, purity statement referencing religious feelings resonate well among religious individuals but are perceived as little convincing by others. GPT-3-generated statements are generally perceived to more convincing than human-generated statements. All the highest ranked statements in terms of convincingness were GPT-3 generated. But the language model struggles with creating novel arguments, human generated statements were more likely to be perceived as novel than GPT-3 generated ones. On the other hand, respondents found novel arguments less convincing and applicable.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
2. Chen, K., Shao, A., Burapacheep, J. and Li, Y. (2022). A critical appraisal of equity in conversational AI: Evidence from auditing GPT-3's dialogues with different publics on climate change and Black Lives Matter. arXiv:2209.13627
3. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P. and Ditto, P.H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. Advances in experimental social psychology, 47, 55-130.
4. Hurst K. and Stern, M.J. (2020). Messaging for environmental action: The role of moral framing and message source. Journal of Environmental Psychology, 68, 101394.
5. Milfont, T.L., Davies, C.L. and Wilson, M.S. (2019). The moral foundations of environmentalism. Social Psychological Bulletin, 14(2),1-25.
6. Moser, S.C. (2016). Reflections on climate change communication research and practice in the second decade of the 21st century: what more is there to say? WIRE's Climate Change, 7(3), 345-369.
7. Syropoulos, S. and Markowitz, E.M. (2021). Perceived responsibility towards future generations and environmental concern: Convergent evidence across multiple outcomes in a large, nationally representative sample. Journal of Environmental Psychology, 76, 101651.
8. Welsch, H. (2020). Moral Foundations and Voluntary Public Good Provision: The Case of Climate Change. Ecological Economics, 175, 106696.
9. Zaval, L., Marowitz, E.M. and Weber, E.U. (2015). How Will I Be Remembered? Conserving the Environment for the Sake of One's Legacy. Psychological Science, 26(2), 231-236.

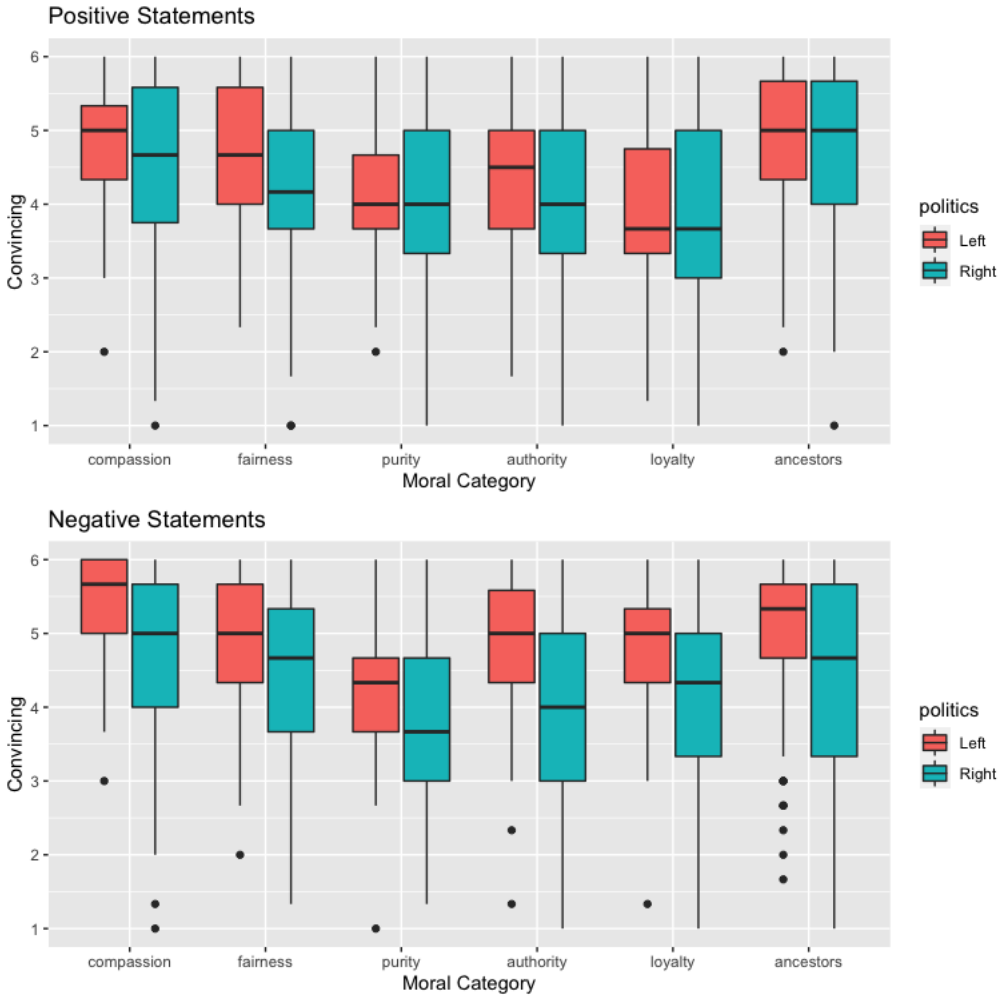

Figure 1. Average convincingness (1 = least convincing to 6 = most convincing) of arguments for climate action based on extended Moral Foundation Theory, comparing politically left- and right-leaning respondents and separated for negative and positively framed argument.