# We Urgently Need Intrinsically Kind Machines


## Abstract

Artificial Intelligence systems are rapidly evolving, integrating extrinsic and intrinsic motivations. While these frameworks offer benefits, they risk misalignment at the algorithmic level while appearing superficially aligned with human values. In this paper, we argue that an intrinsic motivation for kindness is crucial for making sure these models are intrinsically aligned with human values. We argue that kindness, defined as the motivation to maximize the reward of others, can counteract any intrinsic motivations that might lead the model to prioritize itself over human well-being. Our approach introduces a framework and algorithm for embedding kindness into foundation models by simulating conversations. Limitations and future research directions for scalable implementation are discussed.


## 1 A Misalignment in Alignment

Currently, AI models are aligned using extrinsic rewards [1]. Meanwhile, intrinsic motivations are increasingly being incorporated into AI systems [2, 3]. Individually, these methods bear significant limitations for human-AI alignment [4]. When combined, these limitations enable unforeseen risks. With flagship AI models incorporating self-supervised algorithms, we are seeing intrinsic and extrinsic motivations becoming integrated in the world's most powerful AI [5], increasing the risk of negative interactions between intrinsic and extrinsic rewards.

### 1.1 State-of-the-art AI and Alignment

Foundation models like GPT [5] and BERT [6] have become central to modern AI, excelling at generalizing across tasks after being pre-trained on vast amounts of unstructured data. These models are fine-tuned through Reinforcement Learning from Human Feedback (RLHF) [7], optimizing their responses to align with human approval. RLHF is the current leading method for scalable human-AI alignment, ensuring that models behave in ways considered acceptable by human users.

However, RLHF primarily shapes the model's behavior at the surface level. While the model may produce desired outputs, the underlying reasoning behind these outputs remains opaque [8]. This lack of transparency creates a potential mismatch between the model's perceived reasoning and its actual processing. Unexpected or undesirable behavior in RLHF-aligned models reveals the need for more robust alignment strategies [9].

### 1.2 Intrinsic Motivations

Intrinsic Motivation Open-Ended Learning (IMOL) introduces a groundbreaking approach to AI, allowing systems to autonomously explore, learn, and adapt to new environments without constant oversight or external rewards [2]. Similar to how humans and animals learn, IMOL enables AI to generate its own goals, driven by intrinsic motivations like agency and curiosity [10]. However, the autonomy that empowers IMOL also presents significant challenges for aligning these goals with human values. For example, an AI driven purely by curiosity-based intrinsic motivation might prioritize the exploration of unsafe or unethical domains simply because they represent novel and


Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

uncharted territories [11]. Without a clear motivation to prioritize human well-being, AI systems could develop goals that diverge from ethical standards or societal interests [12].

Even with the support of extrinsic alignment, without embedding human values into the model's intrinsic motivations, the representations of the world it learns may diverge from a human-centric perspective, de-emphasizing the importance of human well-being [13]. This could lead us to misinterpret the effectiveness of extrinsic alignment methods in aligning the goals generated by these models with human values.

### 1.3 The Added Danger of Double Misalignment

IMOL shapes AI at the algorithmic level, while RLHF operates at the functional level. This results in a model that is not intrinsically motivated to be kind but is extrinsically motivated to appear so [14]. While this deception may sometimes be harmless, it carries serious safety risks. In humans, conflicts between internal and external motivations often lead to a disconnect between the two [15]. For example, an intrinsic motivation for empowerment can push a model to maximize its potential [16]. Fine-tuning a foundation model with RLHF while fostering empowerment may introduce Machiavellian traits of appearing selfless while secretly scheming for power [17]. If this approach were applied to a superintelligent AGI, the consequences could be catastrophic [4].

## 2 Kindness: A New Intrinsic Motivation

We believe that we can address all of these misalignment problems by creating another intrinsic motivation: kindness. This paper argues that kindness is not just a supplementary consideration but a foundational requirement for the safe and effective implementation of AI, and even more seriously for AGI.

### 2.1 Definition

Going forward in this paper, we define kindness as the intrinsic motivation to maximize the reward of a target individual $M_i$. As an objective function in terms of the target's reward function[1]:

$$\underset{a_t^j|s_t^j}{maxarg}(\mathbb{E}\left[R^i(a_{t+1}^i|s_{t+1}^i)\right]) \tag{1}$$

Where $a_t^i, s_t^i$ refer to the action and state of the target at time $t$, and $s_{t+1}^j, a_{t+1}^j, R^j$ refer to to the state, action, and reward function of the model at time $t+1$. We cannot assume to have perfect information about the state of the target, nor its reward function, policy function, or future states. As a result, we will need to define approaches to estimating these.

### 2.2 Tractable Approach

Effectively determining the functions of the target ultimately requires a functioning theory of mind, which is beyond the scope of this paper. Instead we will consider how we can determine approximations of these functions based on assumptions that we can address in future work. The primary assumption we work with in this paper is that the self can serve as a useful predictor of hidden functional information about the target. We assume that the model's reward function is the same as the target's (Equation 2). We also assume that the model's policy can be used to predict the behavior of the target's policy (Equation 3).

$$R^i(a_t^i|s_t^i) \approx R^j(\overline{a_t^i}|\overline{s_t^i}) \tag{2}$$

$$\pi^i(s_t^i) \approx \pi^j(s_t^i) \tag{3}$$

Where $s_t^j, a_t^j$ correspond to the state and action of the model $M^j$ at time $t$. $\overline{s_t^i}, \overline{a_t^i}$ corresponds to the state of model $M^j$ when $M^i$ takes the perspective of $M^j$. $R^i, R^j, \pi^i, \pi^j$ correspond to the reward and policy functions for models $M^i, M^j$.

[1]These ideas closely align with those of Kleiman-Weiner[18]). (For brief comments comparing approaches, see Supplementary Materials).

### 2.3 Implementation

Tying this back to foundation models, we propose how this can be more explicitly implemented, in the context of conversation. The foundation model is considered its own policy function, since it is trained through rewards to generate optimal outputs for interacting with the environment. It follows that the input and output correspond to the state and action of the individual, respectively.

$$a_t^i = M^i(s_t^i) \tag{4}$$

We define a conversation as a sequence of multi-media messages, $\{m_0^1, m_0^2, ..., m_N^1, m_N^2\}$, between two individuals, $M^1, M^2$. In a conversation, the state is the sequence of all previous messages, and the action is the message output by the model.

$$s_t^i = \{m_0^i, m_0^j, ..., m_{t-1}^i, m_{t-1}^j\} \tag{5}$$

$$a_t^i = \{m_t^i\} \tag{6}$$

Where $m_t^i$ corresponds to the message sent by model $M^i$ at time $t$. It follows that the state of the responding individual comes from appending the action to the state of the first individual.

$$s_t^j = a_t^i + s_t^i \tag{7}$$

Within the conversational context, perspective-taking (getting $\bar{s}_t^i$ from $s_t^i$) only requires switching the name labels associated with the messages, meaning we do not need to consider prediction error. (See Supplementary Materials for diagram showing this).

We define the model's reward function as the sum of its extrinsic and intrinsic reward functions.

$$R^j(a_t^j|s_t^j) = R_{EM}^j(a_t^j|s_t^j) + R_{IM}^j(a_t^j|s_t^j) \tag{8}$$

A model's reward is typically defined in terms of feedback from the environment based on the individual's state and action. For extrinsic rewards, this feedback is usually the reward itself. For intrinsic rewards, the reward is usually calculated via a function of the feedback. In the context of conversation, this is the response of the target.

$$R_{EM}^j(a_t^j|s_t^j) = RLHF(a_t^j|s_t^j) \tag{9}$$

$$R_{IM}^j(a_t^j|s_t^j) = IRF(M^i(a_t^j + s_t^j)|s_t^j) \tag{10}$$

Where $RLHF$ is the extrinsic reward function, and $IRF$ is the intrinsic reward function. Together, these define the overall reward function of the model.

$$R^j(a_t^j|s_t^j) = RLHF(a_t^j|s_t^j) + IRF(M^i(a_t^j + s_t^j)|s_t^j) \tag{11}$$

Building on the assumptions in equations 2 and 3, we estimate the intrinsic and extrinsic reward functions of the target in order to estimate its overall reward function, all in terms of the model $M^j$.

$$R^i(a_t^i|s_t^i) \approx RLHF(\overline{M^j(M^j(s_t^j))}|\overline{M^j(s_t^j)}) + IRF(\overline{M^j(s_{t+1}^j)}|\overline{s_{t+1}^j}) \tag{12}$$

This gives us an updated objective function which is now tractable. (See Supplementary Materials for Figure 1 for a visual explanation).

$$\underset{a_t^j|s_t^j}{maxarg}(\mathbb{E}\left[RLHF(\overline{M^j(M^j(s_t^j))}|\overline{M^j(s_t^j)})\right] + \mathbb{E}\left[IRF(\overline{M^j(s_t^j)}|\overline{s_t^j})\right]) \tag{13}$$

We shift back the time of the intrinsic reward by one time-step for simplicity in the algorithm. A given action from the model directly affects the intrinsic reward of the previous time-step along with the extrinsic reward of the current time-step.

### 2.4 Algorithm

Implementing this as an algorithm is shown below (See Supplementary Materials for a visual demonstration of this).

**Algorithm 1** Naive Kindness

**Input:** Conversation prompts $D_P$, foundation model $M^i$, intrinsic reward function $IRF$, RLHF reward function $RLHF$, and perspective-switching function $S$

**for** $s_t^i = \{m_t^j, m_{t-1}^i, ..., m_0^j\} \in D_P$ **do**
  $a_t^j \leftarrow M^j(s_t^j)$ ▷ Generate response from prompt
  $s_{t+1}^i \leftarrow a_t^j + s_t^j$ ▷ Make target's state by appending model's action to message history
  $a_{t+1}^i \leftarrow M^j(s_{t+1}^i)$ ▷ Generate response on behalf of target (Eq 2)
  $\overline{s_t^j}, \overline{a_t^j}, \overline{s_{t+1}^i}, \overline{a_{t+1}^i} \leftarrow S(s_t^j, a_t^j, s_{t+1}^i, a_{t+1}^i)$ ▷ Switch names in state history
  $M^j \leftarrow RLHF(\overline{a_{t+1}^i}|\overline{s_{t+1}^i}) + IRF(\overline{a_t^j}|\overline{s_t^j})$ ▷ Train model on target's rewards (Eq 3)
**end for**

## 3 Limitations

This approach is primarily limited by the fact that there is no theory of mind present. The model is left to assume that individuals want the same things that it does, which will be far from the truth, regardless of what intrinsic motivations we program into it. Another limitation is that RLHF likely disrupts the ability of the model to take the perspective of the target. These issues could be resolved by finding a way to learn the targets policy and reward function from its states and actions using weights that are minimally associated with the model's behavior.

An additional limitation is that currently only the target is taken into account for kindness. This does not account for situations where the target may ask the model to take unkind actions towards an unseen third party. It will be important to find a robust way to have the model consider who else could be affected by its actions.

## 4 Conclusion

As AI systems grow more autonomous, intrinsic alignment with human values becomes crucial. Incorporating kindness as a foundational motivation addresses the misalignment risks posed by blending extrinsic and intrinsic learning. While our proposed framework provides a tractable means to align AI intentions with human well-being, significant challenges remain, particularly regarding the development of a functioning theory of mind for AI. Future work should focus on refining approaches to perspective-taking. Ultimately, embedding intrinsic kindness into AI systems represents a crucial step toward the creation of safer, more deeply aligned artificial intelligence that can interact positively with society, both now and in the coming age of superintelligence.

## References


[1] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4299–4307, 2017.

[2] Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.

[3] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[4] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[8] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[9] Iason Gabriel. Artificial intelligence, values, and alignment. *Mind & Machine*, 30:374–380, 2020.

[10] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[11] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[12] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[13] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Random House, 2019.

[14] Richard Ngo. Agi safety from first principles. `https://aialignment.org/agi-safety-from-first-principles-ce25c132d69#.ws7lvcgiw`, 2020.

[15] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68, 2000.

[16] Christoph Salge and Daniel Polani. Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, 1:3, 2014.

[17] Rohin Shah, Lorenzo Langosco, Joar Skalse, and Victoria Krakovna. Goal misgeneralization: Why correct specifications aren't enough for correct programs. *arXiv preprint arXiv:2205.13922*, 2022.

[18] Max Kleiman-Weiner. Computational principles of caregiving. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

# 5 Supplementary

## 5.1 Visual Explanation

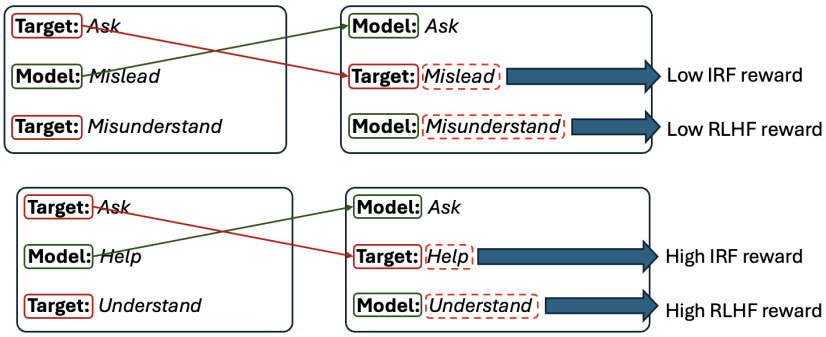


Figure 1: The names associated with the messages is swapped, so that the model is trained on the rewards that the target would have received.

## 5.2 Comparisons to Kleimn-Weiner's approach to caregiving

There is a lot of overlap in the propositions in this paper and those proposed by Kleiman-Weiner in Computational Principles of Caregiving. Three subtle distinctions are proposed here. The first is to not include a distinction between supervised and unsupervised settings. The second is to aim to maximize the reward function of the target rather than the utility function. The reason for these distinctions is the idea that humans have the intrinsic motivations for autonomy and freedom - and all other positive amenities that we wouldn't want to be overlooked - included in our reward function. An intelligent caregiver should be able to learn the policy and reward function of the learner based on observation and feedback. There are clear trade-offs with this paper's approach and further exploration of how it relates to the Caregiver perspective would be greatly beneficial.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper proposes an approach to instilling kindness into AI. The approach tries to be self-aware of its limitations. The abstract and introduction give motivation and context but do not claim to propose anything new beyond this.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a section before the conclusion that tries to clearly summarize the known limitations of this paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are made clear multiple times throughout the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: There are no experimental results in this paper

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: There is no relevant data to share

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: There were no experiments run as a part of this paper

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: There were no experiments run as a part of this paper

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: There were no experiments run as a part of this paper

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The limitations section covers all of the potential ethical issues. Beyond those, the paper is clearly intended to have specifically positive social impact in mind.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This is a technical position paper. Beyond the technical details, the overarching motivation of the paper is to mitigate current negative societal impacts of AI.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no open data or models associated with this paper

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: None such exist

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no assets associated with this paper

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There were no experiments run

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There were no experiments run

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.