Asymmetric control in LLMs: model editing and steering that resists control for unalignment

Summary

Recent efforts in concept level model steering such as Activation Addition or Representation Engineering as well as more granular model editing techniques such as ROME or scrubbing techniques like LEACE are promising approaches towards natural language generation control that is aligned with human values that may skirt some of the problems with Reinforcement Learning From Human Feedback. However there is a big problem with these approaches: they could be equally used by bad actors to unalign models and inject misinformation. In fact the major benchmark, Counterfact, for model editing determines the efficacy of an intervention based on the degree to which we can make models less truthful! Similarly steering methods have demonstrated their efficacy by showing how they can make models less moral or more power seeking (see examples in Representation Engineering paper).

This project involves developing a research direction where control interventions would be ineffective for counterfactual editing or unaligned control but remain effective for factual editing and aligned control. We call this **asymmetric control** since control can only happen in a direction towards alignment with human values not away from it. *If we could achieve this* then editing and steering interventions developed along these lines would be much safer reducing the risk of externalities (safety research being put to bad use). *If we can't achieve this* then we should have empirical evidence saying so and should be much more careful with the development control techniques. Either outcome should provide empirical evidence for or against the orthogonality thesis in prosaic alignment settings.

Not convinced? **Neither are we**. The purpose of this project is to explore *the possibility* of asymmetric control within the context model editing and steering in large language models by developing an empirical setting under which it can be evaluated. We equally want to know whether asymmetric control is a pipedream or a tractable research direction and we'd like to develop a project that can determine this.

The goals of our project is to:

- Conceptualise the setting of prosaic (within current LLMs) asymmetric control **operationally** (meaning we want a framework that yields an empirical assessment).
- Formulate an empirical setting under which **asymmetric control** / control symmetry can be evaluated for model steering and editing
- Measure the current symmetry of various control methods.
- *Attempt* to develop a mechanism for asymmetric control.

- Communicate these findings to both NLP and AI Safety folks so there is clarity on the risks of control.

The non-goals of this project but
are-still-interesting-and-hopefully-our-project-can-provide-foundations-for are:

- Ensuring our conceptualization works for (i) reinforcement learning (ii) agentic systems (iii) non-prosaic AI systems
- Show we can do asymmetric control (we don't know this)

This project is hopefully an interdisciplinary collaboration between folks interested in conceptualising and evaluating asymmetric control in an operational prosaic setting and folks interested in the empirical demonstration and technical development of interventions that may or may not prove it out.

The non-summary

The problem

Current methods on steering and editing are not asymmetric; they (appear to) implement symmetric control. Symmetric control means the edit or steering technique can equally be used to align or unalign a language model. This means that research in this direction is equally beneficial to good and bad actors. This is very concerning, especially since researchers in this subfield are motivated by safety concerns (it's in pretty much every introduction section of these papers) and suggest their works address safety concerns. But we want ***asymmetric control***, meaning we want techniques that work for steering LLMs towards more helpful, honest, and harmless directions but wouldn't even work for steering towards unaligned directions or injecting misinformation.

Solution part 1: Conceptualise

The first thing to do is characterise symmetric control. The fact is we do not know the extent of symmetry in current control methods. This is actually the majority of the project: defining control symmetry and how to evaluate it.

We will only be able to do this with a solid conceptual definition of both symmetric and asymmetric control and an understanding of the risk landscape of this setting, the properties we should evaluate, and the dimensions and settings of control to evaluate.

The outcome of this conceptualization should be an operational framework which helps us develop an evaluation setting for asymmetric control. To this end, we will develop a benchmark and set of evaluation protocols to empirically assess model steering and editing interventions.

Some questions we will need to answer:

- How is this different from RLHF, inoculating bad control in the prompt space?
- What are the reasonable limits or expectations we can have on control symmetry?

- In order to have asymmetric control, what kind of interventions should we expect to develop?

Solution part 2: Empirical assessment

We want to know to what extent current control techniques implement control symmetry. Using the framework developed above, we will run a set of experiments so we can empirically understand the current state of control symmetry.

Solution part 3: Implement control asymmetry

This is the moonshot part of our project. We don't know how we might implement this nor whether it would work. We will attempt a few of these.

A set of baselines that I have identified that are easy to develop but not, i imagine effective, are:

- Prompt engineering for In context knowledge editing
- Identifying an activation vector that resists bad steering but works for good steering
- Train a model to resist bad edits / steering simply by having negative samples, using a control token, constructing a dataset of acceptable and unacceptable steering at the prompt level.

The above are bad for a number of reasons like not being general to all settings, not actually solving the problem of "inoculating" models to bad control, and being ad-hoc.

A set of research directions we could explore here are:

- **In Activation Space**: Understanding the representation or activation landscape behind value-aligned behaviour and unaligned behaviour and attempt to use those insights to fix unaligned control attempts somehow
- **In Weight Space:** Develop an inoculation procedure to train a model to be innoculated against unaligned edits or steering in the weight or activation space, not simply the prompt space.

How does this help with reducing risk?

The primary way this helps risk is that if we can have asymmetric model steering and model editing techniques then we can (1) justify research on the safe continuation of steering and editing research (2) ensure the research won't be effective for bad actors (3) use these techniques to get all the safety benefits of steering and model editing.

A secondary benefit is that while this research explores a very narrow prosaic subset of control, if we can develop a technique of alignment-resistant or asymmetric control, we might be able to work on making this research more general so that it applies/inspires/provides fuel for research in non-prosaic settings like asymmetric control of superintelligence.

An imagined beneficial outcome of the "inoculation" research direction is that model developers could certify their models as "inoculated" against unaligned control.

Some research results we should expect that help with thinking about risk are: what kind of control symmetry currently exists? What does this say about orthogonality, a critical premise for many x-risk arguments, that is currently almost purely a theoretical speculative notion.

The project plan

This is a hypothetical plan that isn't meant to be read as temporally sequential.

- Develop a conceptual definition of alignment-resistant model steering and model editing. The definition will include what does and does not count as alignment-resistant. The definition *must* be operational for empirical NLP research.
- Survey current steering and editing evaluation setups with an eye towards (i) identifying gaps in valid and reliable evaluation (ii) developing a benchmark that is able to evaluate symmetric and asymmetric control as well as desirable types of control we want to allow such as fact correction or morality and ones we want resistance too such as injecting misinformation or steering towards power seeking.
- Develop a benchmark that would be empirically satisfying to demonstrate asymmetric control
- Develop the stupidest simplest baseline intervention for doing asymmetric control
- Develop a comprehensive set of experiments that demonstrate (or not) asymmetric control of common steering and editing methods.
- Get feedback on that design and execute it
- Perform post-hoc analysis such as manual error analysis or additional control experiments to explore the results
- Attempt to develop asymmetric control techniques

Scope limitations: What we might not do

Develop novel interventions **beyond** the stupidest simplest baseline we can think of. If the team is large enough we can hopefully have at least one member dedicated to developing asymmetric control methods - they may or may not discover a tractable implementation and that is ok and good information.

Do both model editing and model steering - we might just select one.

Scope limitations: What we likely won't do

1. Unify steering and editing conceptually (this seems hard and a separate project)
2. Unify fact injection / updating and fact / concept deletion/ scrubbing (this seems hard and a separate project)
3. Develop anything other than an *operational* definition of asymmetric control that works for steering and editing within LLMs in the context of NLP

Output

The primary expected output is a publication that is of academic quality to be published at an appropriate AI Safety or NLP venue such as workshop on AI Safety at Neurips. The publication will at minimum contain a conceptual specification of what asymmetric control looks like technically, how to assess this (evaluation protocol), and empirical results on how current techniques do or do not implement asymmetric control. A more ambitious project will also propose novel techniques for asymmetric control.

While the primary output will hopefully provide inspiration and empirical results for the steering and model editing communities. We also want a secondary output in the form of an AI alignment / Less Wrong post and/or other that communicates our findings to folks who might be able to work on asymmetric control in other contexts such as a more theoretical and conceptual setting.

Risks and downsides

One risk is that if a model is able to resist control this seems like a bad thing even if it's resisting control in a good direction. There a number of reasons why we don't think externalities of this research increase lack of control: (1) we are looking for methods that can't be used to resist aligned changes (2) methods that could symmetrically provide resistance simply reduce to the current paradigm of bad actors and good actors equally benefiting from this research (but we will not encourage this direction) (3) resistance of control within an agentic framework versus resistance of control in a prosaic language modelling direction seem very different: there doesn't seem like any way research in this direction could be ported to agentic modes of control research for good or bad.

The major downside is if our work has disappointing results showing alignment-resistant control doesn't initially appear tractable. However, this could have a good effect of encouraging a discussion around not continuing to pursue symmetric editing and steering methods due to their risks.

Another downside is that if alignment-resistant methods do end up working, this could lead to people assuming these methods are sufficient for AI safety when they are not. Using these methods to ensure safety but failing to understand this is only one small part of the AI safety solution is a potential risk.

Finally, deceptive control might still be at play here: first, we think theoretical guarantees of asymmetric control are necessary to be able to say how effective control asymmetry is across distributions. Without this, there is always a chance that there is a place out-of-distribution that control symmetry might exist. In this sense we might be deceiving ourselves. More insidious, there is the possibility of deceptive alignment where control appears asymmetric in order to hide the possibility of control in unaligned directions.

We don’t see any part of this project as presenting an infohazard or encouraging capabilities progress. While steering can somewhat improve capabilities on some tasks, they are generally ineffective for increasing capabilities and alignment-resistant steering would only make them less effective for capability increase.

Acknowledgements

Activation Addition and Representation Engineering papers were particularly influential.

Team

Team size

A small team of 3-5 people passionate about the project is sufficient. Ideally there will be a balance of people who would like to think formally and conceptually about asymmetric control or impact, safety risks, and safety assessment as well as people who like doing empirical NLP work using code and models and like to think about conceptually and theoretically and would like to bridge the gap to thinking about technical alignment interventions.

We are open to expanding our team size if we get enough interest so that we can cover four separate directions:

(1) Conceptualization
(2) Evaluation and Assessment
(3) Empirical Experimentation
(4) Novel asymmetric control technique development

Research Lead
Domenic Rosati
domenic.rosati@dal.ca

I have been an NLP researcher since 2016 mostly focusing on NLP applied to scientific texts. Since 2020, I have been working on safety in natural language generation in biomedical and scientific settings at scite.ai. I recently started my PhD at Dalhousie with Hassan Sajjad and Frank Rudzicz on technical approaches to alignment where my focus is on Model Editing and Steering as safety interventions. I have three papers on self-consistency for AI Safety (BlackboxNLP@EMNLP, Main Track AAAI, AiSafety@Neurips), one of which won the best paper award at AI Safety workshop at Neurips and another forthcoming work on model editing evaluation techniques. I have a cute whippet and two kids and I live in Halifax and I like reading science fiction. My research vibe is easy-going, intuitive, exploratory, and insight driven.

I will spend 20 hours per week on the project + more for mentorship, team coordination, meetings, writing, and other.

Team Coordinator

I would prefer someone to volunteer to be the team coordinator as I am disorganised and chaotic by nature.

Skill requirements

We'd like a project that has a balance of skills within the AI safety domain.
The following four areas of skill are of interest. Members of the group only need proficiency in at least one of these areas, we will recruit members for a balance of these.

Since this is a mentoring opportunity, proficiency here is defined as enough competency with the methods in the area to work individually with direction.

The code of the project will be all written in python but only members contributing to the empirical work will need proficiency here.

Conceptual Alignment / AI Safety

- Basic understanding and familiarity of the major concepts in AI/Value Alignment literature
- Ability to formalise conceptual work (such as using first-order logic for definitions and others)
- Passion for thinking conceptually through motivation, experimental proof, theoretical proof of proposed interventions
- Basic familiarity with some theoretical methods and tools such as first-order logic, upper and lower bound analysis, proofs, philosophical analysis

Risk Assessment / Evaluation

- Basic understanding and familiarity of the research landscape in AI Risk
- Basic understanding and familiarity (mainly technical) AI risk assessment and evaluation
- Creative thinking around experimental design for evaluation

Deep learning

- Understanding of Deep Learning
- Comfort and competency with design, training, and evaluation of neural networks
- Creative thinking around technical approaches to alignment
- Comfort in doing empirical research using neural networks

NLP

- Understanding of Large Language Models and Transformers
- Understanding of Interpretability and Representation Probing techniques
- Comfort in modelling and evaluation of NLP techniques