Research Outline on Mechanistic Interpretability)

Name: *Abid Jeem*

March 12, 2025

1. **Introduction and Motivation**

   Mechanistic Interpretability (MechInterp) seeks to reverse-engineer neural networks by uncovering and understanding their internal processes, much like analyzing the inner workings of a machine. In contrast, Explainable AI (EAI) focuses on producing human-friendly explanations for the outputs of AI systems without necessarily exposing the underlying neural mechanisms. This research aims to explore the technical feasibility of MechInterp and investigate its potential impact in the field of Human-Computer Interaction (HCI) by enhancing trust, usability, and the overall human-AI experience.

2. **Distinguishing MechInterp and Explainable AI (EAI)**

   - **Mechanistic Interpretability (MechInterp):** Involves dissecting neural networks to reveal how specific inputs propagate through the model and produce outputs. It is a bottom-up approach that explains the behavior by directly interpreting internal neural representations and dynamics.
   - **Explainable AI (EAI):** Focuses on providing post-hoc, high-level explanations that make the decisions of AI systems understandable to users. It often uses approximations or surrogate models to communicate complex model behavior without delving into the exact internal mechanisms.

3. **Practical Impact in HCI**

   - **Enhancing User Trust:** By exposing how decisions are made at a mechanistic level, systems can offer deeper transparency, thereby fostering increased trust among users.
   - **Improving Interface Design:** Insights from MechInterp can be used to design interfaces that adapt dynamically to user needs and clearly present how inputs are processed, reducing confusion and promoting intuitive interactions.
   - **Personalized User Experiences:** Understanding internal neural patterns can help in tailoring AI behavior to match individual user expectations, leading to more effective and personalized interfaces.

4. **Existing Tools and Libraries**

   Though MechInterp is an evolving field, several existing tools and libraries facilitate neural network interpretability:

   - **Lucid:** A suite of infrastructure and tools for visualizing and understanding neural network internals.
   - **Captum:** A PyTorch library offering state-of-the-art interpretability algorithms, allowing researchers to analyze and attribute predictions to input features.
   - **Distill:** An interactive journal that often publishes clear, interactive articles on machine learning interpretability techniques.

5. **Canonical Papers (Post-2020) and Key Figures**

   - **Canonical Papers:** Research papers exploring mechanistic interpretability have emerged after 2020, including works by Chris Olah and collaborators that delve into neural circuit analysis and visualization. (For example, see recent publications on interpretable bases and neural circuit dissection.)
   - **Key Dignitaries:**
     - **Chris Olah:** Widely recognized for pioneering research in mechanistic interpretability.
     - **Shumin Zhai:** Known for contributions to HCI and the integration of human performance models with AI.
     - **Himabindu Lakkaraju:** Notable for work in interpretable and fair machine learning models bridging AI and HCI.

6. **Research Plan and Outline**

   The proposed research plan will follow these stages:

   (a) **Literature Review:** Conduct a comprehensive survey of current research on both MechInterp and EAI, identifying strengths, limitations, and open questions.

   (b) **Gap Analysis:** Determine how mechanistic interpretability can address key challenges in HCI, such as enhancing transparency, trust, and user-centric design.

   (c) **Prototype Development:** Develop experimental prototypes that integrate MechInterp insights into HCI contexts. This could involve:

   - Building visual interfaces that expose internal neural activations.
   - Creating interactive dashboards that allow users to explore model behavior in real time.

   (d) **User Testing and Evaluation:** Design and conduct user studies to assess the impact of these prototypes on user trust, usability, and understanding. Metrics might include task performance, qualitative feedback, and usability scores.

(e) **Iterative Refinement:** Based on feedback from initial studies, refine both the interpretability methods and the HCI interfaces to better meet user needs.

(f) **Dissemination:** Prepare scholarly articles and presentations that document the methods, prototypes, and user study results, contributing to the academic discourse in both AI interpretability and HCI.

7. Conclusion

This research has the potential to bridge the gap between low-level neural analysis (Mech-Interp) and user-centric design (EAI in HCI). By systematically exploring the technical foundations, practical applications, and human implications of mechanistic interpretability, the study aims to pave the way for more transparent, accountable, and engaging AI systems.