Lab 06: Museum Exhibit Design - Neuroscience and AI Connections

Course: Neuroscience as a Model for AI

Professor Patricia McManus

Benjamin Bui-Dang W211756294

Date: April 08, 2025

Section 1: Neuroscience Concept Definition

A neuroscience approach called predictive coding views the brain as a prediction machine that is always creating models of the outside world to foresee incoming sensory data. This theory holds that the brain does not passively interpret sensory inputs; instead, it maintains hierarchical generative models that predict them. These internal models are updated in response to "prediction mistakes" that arise when sensory input and predictions diverge. Multiple brain processing levels are involved in this process, with higher levels producing predictions about activity at lower levels.

Many perceptual phenomena are explained by this theory, such as how we can perceive whole objects even when we have incomplete sensory information (filling in blind patches or obstructed objects) and why we occasionally encounter sensory illusions when our preconceived notions take precedence over unclear sensory inputs. A major change from considering perception as a bottom-up process to seeing it as a dynamic interaction between top-down predictions and bottom-up sensory data is represented by predictive coding.

Section 2: Connection to Artificial Intelligence

Predictive coding principles have profoundly influenced modern AI architectures, particularly in computer vision, natural language processing, and generative models. Several AI systems explicitly incorporate this neuroscience concept:

Predictive Coding Networks (PCNs): These neural networks implement hierarchical probabilistic predictions, with each layer predicting the activity of layers below it. PCNs have demonstrated human-like visual processing capabilities, including robustness to noise, illusion susceptibility, and fill-in completion of partial information.

Variational Autoencoders (VAEs): These generative models learn latent representations of data by balancing reconstruction accuracy against prior expectations, directly implementing the prediction-error minimization principle from predictive coding theory. VAEs excel at generating new examples that fit learned categories while maintaining uncertainty in their predictions.

Large Language Models with Attention: While not explicitly designed as predictive coding systems, transformer-based models like GPT-4 and Claude effectively implement prediction mechanisms by learning to anticipate subsequent tokens based on context. Their attention mechanisms create dynamic hierarchical models that update predictions based on contextual information.

Deep Boltzmann Machines: These bidirectional neural networks implement generative models that capture the statistical structure of input data, allowing them to both recognize patterns and generate expected outputs, mirroring the brain's predictive capabilities.

The predictive coding framework explains why AI systems that incorporate hierarchical generative models demonstrate more robust, human-like intelligence compared to purely discriminative approaches. These systems can handle uncertainty, form abstractions, and generalize to novel situations more effectively by learning to predict rather than simply classify.

Section 3: Exhibit Proposal

Exhibit Title: "Expecting Minds: How Prediction Powers Perception and AI"

Format: Interactive mixed-reality installation with digital displays, physical manipulatives, and augmented reality components.

Content Summary: The exhibit will create an immersive exploration of predictive coding through three interconnected zones:

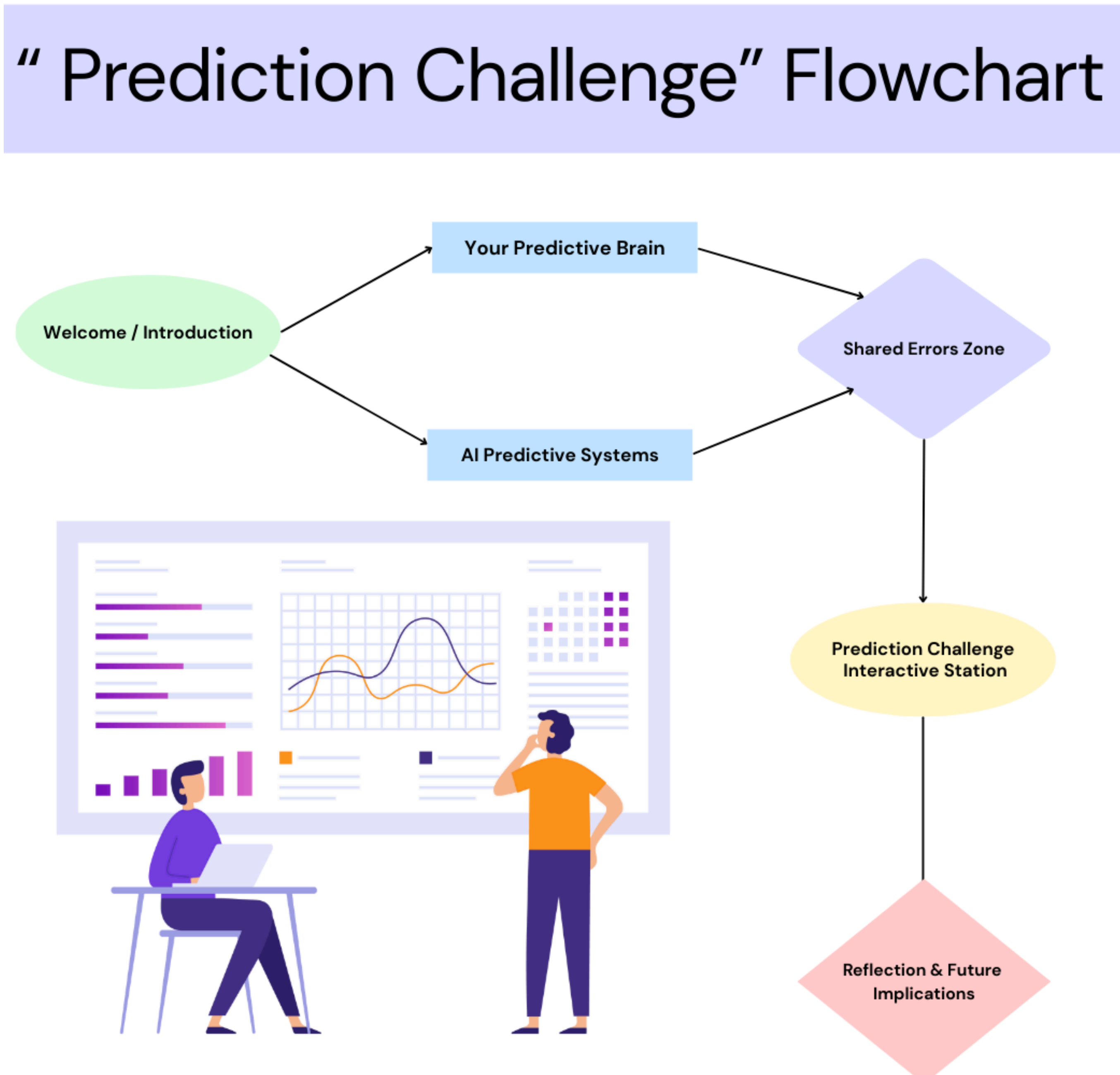
1. **"Your Predictive Brain" Zone:**
    - Visual illusions displayed on large screens demonstrate how prior expectations shape perception
    - Audio stations where visitors experience phonemic restoration (hearing complete words despite missing sounds)
    - Tactile puzzles illustrating how the brain completes missing sensory information
    - Educational panels explaining the neuroscience of hierarchical prediction in cortical processing
2. **"AI Predictions" Zone:**
    - Transparent displays showing real-time visualizations of predictive AI systems processing sensory information

    - Video demonstrations comparing traditional bottom-up AI approaches with predictive coding networks
    - Case studies of AI applications using predictive principles (medical imaging, robotic perception, etc.)
    - Educational content explaining similarities between hierarchical neural networks and brain organization
3. **"Shared Errors" Zone:**
    - Side-by-side demonstrations of situations where both human brains and AI systems make similar predictive errors
    - Exploration of how both systems can be "fooled" by manipulating expectations
    - Discussion of ethical implications when AI systems make prediction errors

Interactive Component: The centerpiece of the exhibit will be the "Prediction Challenge" station, where visitors engage with both a neural network and their predictive capabilities:

1. Visitors don augmented reality glasses showing a scene with multiple objects
2. The system gradually degrades or occludes parts of the scene
3. Visitors are asked to predict what objects are present or what will happen next
4. Simultaneously, an AI system makes its predictions about the same scene
5. Results compare the visitor's predictions with the AI's predictions, highlighting similarities in error patterns
6. The experience concludes with a personalized visualization showing how the visitor's brain and the AI system used similar predictive mechanisms

Visual Diagram: Exhibit Layout



Section 4: Reflection and Justification

Because predictive coding signifies a paradigm change in our understanding of artificial intelligence and human cognition, I chose it as the exhibit's main topic. Predictive coding provides a unifying framework that explains a variety of occurrences across many sensory modalities, in contrast to more specific brain mechanisms. Because it offers an obvious metaphor—the brain as a prediction machine—that visitors can connect to through common experiences, this idea is very potent for public education.

By bringing to light the typically hidden prediction processes that take place in both brains and AI systems, the exhibit improves public awareness. The exhibit fosters an embodied awareness that transcends intellectual explanations by letting visitors experience both AI forecasts and their own prediction processes. By demonstrating how AI's intelligent behaviors stem from concepts akin to human cognition and emphasizing the complexity of our neurological processes, this method demystifies AI.

The practical implications highlighted by this exhibit are significant. As AI systems increasingly adopt predictive coding principles, they become more robust, adaptable, and aligned with human expectations. The exhibit illustrates how this neuroscience-inspired approach is leading to AI systems that can operate in uncertain environments, anticipate needs, and interact more naturally with humans. It also raises important ethical considerations about how predictive AI systems might amplify certain biases or create new forms of manipulation if predictions are optimized for engagement rather than accuracy.

According to the exhibit, further incorporation of predictive coding concepts into AI architectures may result in systems that are better able to reason causally, learn from sparse data more effectively, and better reflect human values—all of which are important areas of contemporary AI research. The display urges visitors to recognize the amazing powers of the human brain as well as the potential for neuroscience to inform more human-compatible artificial intelligence by relating these technological possibilities to their biological inspiration.

References

Anderson, M., & Chemero, T. (2023). Predictive processing and the free energy principle: A comprehensive review. *Neuroscience & Biobehavioral Reviews, 145*, Article 104913. https://doi.org/10.1016/j.neubiorev.2023.104913

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138. https://doi.org/10.1038/nrn2787

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 35*(2), 209-223. https://doi.org/10.1111/mila.12281

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence, 2*(11), 675-683. https://doi.org/10.1038/s42256-020-00222-y

Millidge, B., Seth, A., & Buckley, C. L. (2022). Predictive coding: Towards a future of deep learning beyond backpropagation? *Neural Computation, 34*(4), 753-787. https://doi.org/10.1162/neco_a_01454

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79-87. https://doi.org/10.1038/4580

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 216*(1205), 427-459. https://doi.org/10.1098/rspb.1982.0085

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences, 23*(3), 235-250. https://doi.org/10.1016/j.tics.2018.12.005

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356-365. https://doi.org/10.1038/nn.4244