Mirror-Neuron Patterns And Self/Other Foundations for AI Alignment

Robyn Wyrick
Department of Computer Science
University of Bath, United Kingdom

wyrickrv@deepalignment.ai

October 2025

Abstract

As artificial intelligence (AI) advances toward superhuman capabilities, aligning these systems with human values becomes critically important. Current alignment strategies rely on externally specified constraints that may prove insufficient against future super-intelligent AI capable of circumventing top-down rules.

This dissertation explores embedding ethical principles in AI by investigating whether artificial neural networks (ANNs) can develop patterns analogous to biological mirror neurons — cells that activate both when performing and observing actions. Mirror neurons play a crucial role in empathy, imitation, and social cognition in humans. The research addresses: (1) Can simple ANNs develop mirror neuron patterns? and (2) How might these patterns contribute to ethical AI systems?

Using a novel "Frog and Toad" game framework designed to promote cooperative behaviors, we identified conditions for mirror neuron pattern emergence, evaluated their impact on decision-making, developed the Checkpoint Mirror Neuron Index (CMNI) to quantify activation consistency, and proposed a theoretical framework for further research.

Findings indicate that appropriately scaled model capacities and self/other attention foster shared neural representations in ANNs similar to biological mirror neurons. This suggests that intrinsic motivations, modeled through mirror neuron patterns, offer a viable pathway to deeper ethical alignment in AI, enhancing capacity for empathetic and cooperative behavior. The findings have implications for both AI

development and broader fields: they contribute to creating AI systems inherently aligned with human values while also providing valuable insights for neuroscience and ethics research.

Keywords: AI alignment, artificial neural networks, cooperation, empathy, intrinsic motivation, mirror neurons

This work was submitted in partial fulfillment of the requirements for the degree of Master of Science in Artificial Intelligence at the University of Bath, December 2024.

Contents

1 Introduction

1.1 Problem Description

As artificial intelligence (AI) rapidly advances superhuman capabilities, the risks of misalignment compound, from amplifying societal biases and inequalities to existential threats that could jeopardize humanity's existence [1–4]. Ensuring that AI systems truly internalize ethical values then, becomes a central challenge.

Current strategies, including value alignment protocols and reinforcement learning from human or AI feedback, provide essential safeguards but rely heavily on external, and rule-based controls. Although these methods are largely effective for managing today's AI systems, they may be insufficient for future super-intelligent AI [2]. Such systems could strategically feign ethical compliance while pursuing harmful or catastrophic objectives [1]. To prevent this, we must look beyond rules and rubrics and foster intrinsic motivations for ethical behavior, embedding foundational principles into the AI's cognitive architecture.

1.2 Intrinsic Motivations Through Mirror Neurons

This dissertation explores whether artificial neural networks (ANNs) can develop patterns analogous to biological mirror neurons, which in humans underlie empathy and social cognition [5]. Mirror neurons fire both when an individual performs an action and when observing that action performed by another [6]. This dual responsiveness contributes to understanding, emotional resonance, and prosocial behavior.

This research seeks to address two pivotal questions::

1. Can simple artificial neural networks develop mirror neuron patterns?
2. How might such patterns contribute to training ethics within AI systems?

If ANNs can foster intrinsic motivations akin to human empathy, this may offer a pathway toward deeper ethical alignment as AI capabilities grow.

1.3 Scope and Objectives

We introduce a controlled experimental framework centered on the *Frog and Toad* game, a minimal environment designed to isolate cooperative behaviors and shared representations. The key objectives are:

1. Determine if and under which conditions mirror neuron patterns emerge in ANNs, focusing on model capacity, game complexity, and the necessity for generalization.
2. Investigate how these patterns influence decision-making, especially regarding self-preservation and prosocial actions.
3. Develop the Checkpoint Mirror Neuron Index (CMNI) to quantify mirror neuron patterns by comparing neural activations across key scenarios.
4. Propose a theoretical framework to explain the emergence of mirror neuron patterns, incorporating concepts like neuronal economy, mutual dependency, and the **Veil of Ignorance**.

This study provides quantifiable evidence that simple ANNs can form shared self/other representations similar to biological mirror neurons. It shows how these patterns support both self-preservation and prosocial behaviors, suggesting a route toward integrating empathy-like processes as intrinsic ethical anchors in AI systems.

2 Literature Review

2.1 Introduction

To explore the research questions posed in the Introduction, this literature review surveys interdisciplinary findings on empathy, social cognition, and alignment strategies. We aim to **trace a link from mirror neuron patterns to empathy, prosocial behavior, and ethics**, guiding future research on their potential contributions to AI ethics. Building on de Waal's Russian Doll model 2008, which positions biological mirror neurons and affective empathy as the foundation of perspective-taking and altruism, we propose that AI systems might similarly develop empathy through the mathematical principles underlying mirror neuron patterns. While unresolved

scientific questions prevent establishing a complete, connected throughline, this paper takes a foundational step: identifying mirror neuron patterns in ANNs.

Mirror Neurons → Affective Empathy → Cognitive Empathy → Prosocial Behavior → AI Ethics

Figure 1: Conceptual Throughline from Mirror Neurons to AI Ethics

2.2 About Mirror Neurons

Mirror neurons, first identified in macaque monkeys in the early 1990s [7], fire both when an individual performs an action and when observing another perform that action [8, 9]. In humans, functional imaging and TMS studies confirm that equivalent mirroring mechanisms exist, contributing to action understanding and empathy [10, 11].

These systems not only match observed actions but also encode goals, emotional states, and perspective [12–14]. Mirror neuron activity has been found in various species, suggesting a general mechanism for interpreting others' actions, emotional expressions, and intentions [15–17].

2.3 Mirror Neurons in Empathy and Social Cognition

Mirror neurons have been extensively studied for their role in empathy, imitation, and social cognition [11, 13, 18]. In humans, mirror neuron systems (MNS) are believed to contribute to understanding others' actions and emotions, forming the basis for empathetic responses. These MNS facilitate the **rapid, automatic and unconscious activation** of neural representations in the observer similar to those perceived in the subject, known as the perception-action mechanism (PAM) [5]. This mechanism is fundamental to emotional contagion, where the observer's emotional state mirrors that of the observed individual.

Several models explain how mirror neurons contribute to empathy:

- **Embodied Simulation**: Ferrari and Gallese 2007 propose that mirror neuron systems, along with other mirroring neural clusters, constitute the neural basis of intersubjectivity. Embodied simulation allows individuals to internally simulate others' actions and emotions, facilitating understanding without conscious effort.
- **Russian Doll Model**: de Waal's 2008 Russian Doll model suggests that higher cognitive levels of empathy build upon basic, hard-wired

processes like emotional contagion. This layered model reflects an evolutionary progression from simple to complex forms of empathy, enabling quick and automatic responses essential for social interactions.

- **Dual Route Model**: Yu and Chou 2018 introduce a dual route model distinguishing between a fast, automatic "lower route" associated with affective empathy and a slower, deliberate "higher route" associated with cognitive empathy.

2.4 From Empathy to Ethics

The relationship between mirror neurons and ethics stems from their role in empathy and social understanding. Decety and Cowell 2014 argue that empathy, emotional sharing, empathic concern, and perspective-taking play pivotal roles in moral reasoning by motivating care for others. Empathic concern, in particular, extends beyond the passive experience of another's pain, driving altruistic behavior aimed at improving their condition [8, 22].

The philosopher John Rawls' **Veil of Ignorance** [23] offers a parallel framework in moral philosophy. Rawls proposed that fair principles arise when individuals are uncertain of their own role or position in a given scenario. Recent empirical studies by Weidinger et al. 2023 demonstrate that similar conditions – uncertainty about self and other – promote fairness-based reasoning and impartial decision-making. This aligns closely with the cognitive processes underpinning empathy, where reduced self/other differentiation fosters mutual understanding and cooperation.

However, empathy does not always lead to ethical actions. Affective empathy can result in bias, in-group favoritism, and even self-protective behaviors [25]. The visceral experience of another's pain can overwhelm an individual, leading to distress-avoidance or actions aimed at reducing one's own discomfort rather than helping the other [5]. On the other hand, cognitive empathy, which involves perspective-taking and the ability to understand another's condition, can help mitigate these pitfalls. Studies show that perspective-taking enables more impartial reasoning and fosters ethical decision-making, particularly in situations where fairness and long-term outcomes must be prioritized over emotional immediacy [22, 24]. By integrating these dimensions, empathy can foster fairness, cooperation, and impartiality—even in complex, uncertain, or emotionally charged scenarios.

2.5 Empathy in Artificial Intelligence

Integrating empathy into AI presents profound challenges, particularly concerning the machine's capacity for subjective experience. However, whether an AI truly has subjective experience or merely seems to have subjective experience, its behavior ultimately reflects the encoded representations within its neural network. While current AI can simulate cognitive empathy via rules and learned patterns, they may fail to sufficiently internalize empathy for ethical reasoning, reducing ethical behavior to a tactical facade [1]. This limitation becomes even more significant when considering future superintelligent AI, whose strategic and operational abilities may surpass human comprehension [2]. Without a model of affective empathy, deeply embedded through mirror neuron-like mechanisms, ethical principles in such advanced systems could remain performative, masking potentially harmful objectives. This paper argues that for AI to truly align with human values, ethical reasoning must be deeply embedded and authentic. Affective empathy, therefore, emerges as a key factor for aligning ethical AI beyond superficial imitation.

2.5.1 Shared Representations and Neuronal Economy

When models are excessively large or unconstrained, they can allocate resources to memorize individual states, bypassing the need for shared representations [26, 27]. Conversely, appropriately scaled ANNs, learn reusable patterns that span multiple scenarios, requiring the network to economize resources and avoid overfitting to specific conditions [28, 29]. This dynamic, which we term **neuronal economy**, is a vital precursor to the empathic-like behaviors explored in this research.

2.5.2 Agent Dependency and the Veil of Ignorance

Empathy also depends on the self/other relationship. Multi-agent tasks and game-theoretic scenarios show that **agent dependency** – shared and dependent outcomes and rewards – encourages cooperative behaviors [30, 31].

Agent dependency's role in empathy is reinforced by limiting **self/other differentiation**, akin to the **Veil of Ignorance** framework in moral philosophy [23], where uncertainty about roles and outcomes fosters impartial decision-making. For example, studies using the Veil of Ignorance show that role uncertainty consistently promotes fairness-based reasoning and cooperative principles, both in theoretical models and experimental AI applications

[24]. These findings highlight how shared interdependencies and uncertain roles encourage strategies that prioritize collective welfare over individual gain, a dynamic critical for designing cooperative AI systems.

2.5.3 Computational Approaches to Affective Empathy

One approach to embedding empathy in AI is the use of computational models of affective empathy. For example, the "Brain-Inspired Affective Empathy Computational Model" integrates the Free Energy Principle to simulate pain and employs a spiking neural network to mimic the human mirror neuron system [32]. While still rudimentary, such approaches hint that embedding empathy-related computations into AI can enhance trust, transparency, and altruistic response

2.5.4 Human-Centered Applications and AI-Assisted Empathy

AI tools have been developed to augment human empathy in healthcare, mental health support, and human-centered design [33–35]. Although these methods primarily employ cognitive empathy and externally specified constraints, they illustrate the potential impact of empathy-oriented approaches. Extending these frameworks to include affective empathy modeled on mirror neuron dynamics – emotional contagion, fast, unconscious internal simulation ([5, 20]) – could yield deeper alignment with human values.

2.5.5 Embodied AI and Homeostasis

Researchers exploring empathetic AI propose integrating embodiment with homeostatic mechanisms – internal processes that maintain stable conditions within an agent despite external changes. Sitti 2021 emphasizes the role of physical intelligence (PI), where an agent's capabilities arise not only from computation but also from the properties of its body. Combined with homeostasis, PI could enable artificial agents to navigate complex environments, much like organisms do, providing a foundation for empathetic behaviors.

Similarly, Man and Damasio 2019 argue that machines incorporating homeostatic principles – ensuring their "virtual bodies" remain within a viable range – gain a form of vulnerability and self-preservation similar to living beings. By striving to maintain internal balance, these systems may adapt, behave intelligently, and potentially express empathetic responses.

2.6 AI Safety, Governance, and Alignment

As AI systems become more capable, ensuring that their goals and behaviors align with human values is increasingly critical to prevent harmful outcomes [38]. Misaligned AI systems pose significant risks, including unintended harm due to poorly specified objectives, unforeseen interactions with their environment, and exploitation of loopholes in their reward functions – known as reward hacking [39]. These challenges highlight the need for robust AI safety measures, governance frameworks, and alignment strategies.

2.6.1 Risks and Challenges in AI Safety

Technical and Socio-Technical Risks AI systems may fail to generalize to new environments, leading to unpredictable or harmful behaviors [39]. Misaligned AI can exhibit goal misgeneralization, feedback-induced misalignment, power-seeking tendencies, untruthful outputs, and deceptive alignment, highlighting the need to embed human values and ethics into AI design [40, 41].

Malicious Use and Global Risks AI may be used in cyberattacks, enhancing physical attacks, and in facilitating political manipulation through surveillance and disinformation [42].

Bias, Fairness, and Societal Harms AI systems often inherit and amplify biases from training data, as demonstrated by higher error rates for darker-skinned females in facial recognition systems compared to lighter-skinned males [4]. Broader societal harms include disinformation, labor market disruptions, and biased or inaccurate outputs in sensitive domains like healthcare. The World Health Organization (WHO) (2024) warns of risks from large language models (LLMs), including automation bias, skills degradation, cybersecurity threats, and challenges in maintaining informed consent in clinical settings.

AI Risks Even When Aligned The risks of Artificial General Intelligence (AGI) misalignment have been extensively documented [38, 43, 44], but Friederich [45] argues that even *successfully aligned* AGI — systems that reliably do what their operators want — poses catastrophic risks through power concentration. When AGI capabilities vastly exceed human intelligence, intent alignment effectively grants near-absolute power to whoever controls

the system, creating pathways to stable totalitarianism or military catastrophe. Friederich proposes that liberal democracies should instead pursue "unaligned symbiotic AGI" developed as an intergenerational social project, where AGI is not subservient to operators but integrated into democratic institutions.

Existential Risk and the AI Doom Debate A growing number of scholars have raised the potential for existential risk from superintelligent AI systems that could surpass human cognitive capabilities [43, 44, 46, 47]. Proponents of the "AI doom" scenario warn that advanced agents with open-ended optimization or recursive self-improvement loops might rapidly circumvent safety measures, leading to irreversible catastrophic outcomes for humanity. Critics argue that these concerns remain speculative given current narrow AI capabilities, and point to humanity's history of navigating other transformational technologies without global catastrophe [38, 48]. Nonetheless, the urgency of these concerns was underscored in 2023 by a widely publicized open letter, signed by some of the world's most prominent technologists and AI leaders, calling for a six-month pause on the development of advanced AI systems to address potential catastrophic risks. The letter described an "out-of-control race" among AI labs to create increasingly powerful systems without sufficient understanding, predictability, or oversight, emphasizing the need for deliberate planning and management to mitigate existential threats [49].

2.6.2 Governance Frameworks and Alignment Strategies

Addressing the risks associated with AI requires a robust ethical governance framework that prioritizes transparency, accountability, safety, and robustness [41, 50]. It requires coordinated international efforts to harmonize regulations, ensuring that AI practices are aligned across borders [3]. Stuart Russell emphasizes that AI systems should maximize human preferences while maintaining uncertainty about these preferences [38].

Below we include some of the leading AI alignment techniques:

- **Interpretability Methods**: Techniques such as LIME, SHAP, saliency maps, and attention mechanisms provide insights into AI decision-making [51]. Frameworks like IBM's AI Explainability 360 and DARPA's XAI program integrate these methods into cohesive systems [52].
- **Assurance Methods**: Safety evaluations, red teaming, and formal

verification ensure AI systems operate predictably and align with human values [2, 40–42].

- **Adversarial Training**: Adversarial examples and robust optimization strengthen AI resilience against manipulative or unexpected inputs [53].
- **Cooperative Training Methods**: Cooperative Inverse Reinforcement Learning (CIRL) and human-AI collaboration promote alignment by fostering collaborative decision-making [38, 53, 54].
- **Reinforcement Techniques**: Reinforcement Learning from Human Feedback (RLHF), Recursive Reward Modeling (RRM), and Reinforcement Learning from AI Feedback (RLAIF) refine AI behavior through iterative feedback [2, 39, 53]. A notable variant is **Constitutional AI** [55], where a model self-critiques and revises outputs based on an explicit, externally crafted constitution, and then uses RLAIF to enforce alignment with those principles.

Interpretability methods are broadly applicable and provide critical insights for both rule-based controls and potential intrinsic mechanisms. However, most alignment strategies beyond interpretability depend on top-down external controls. Huang et al. [56], raise concerns that current value alignment approaches – including RLHF and Constitutional AI – concentrate power in the hands of developers while undermining users' moral and epistemic agency. More centrally for this work, these strategies fail to address **intrinsic motivations**: the internal value structures that determine whether a model genuinely internalizes ethical principles or merely performs compliance [57].

2.7 Summary

This literature review establishes that mirror neuron systems support empathy and social cognition, while empathy in turn guides moral judgments and prosocial behavior. We explored how affective empathy serves as the foundation for our initial understanding of others – facilitating fast emotional contagion and the unconscious perception-action mechanism. We observed that the pathway from empathy to fairness or altruism is not guaranteed, as biases and emotional overload can skew moral actions. By integrating the perspective-taking and impartial reasoning of cognitive empathy, we mitigate these challenges. Concepts like Rawls' *Veil of Ignorance* and the recognition of *agent dependency* demonstrate that uncertainty about one's role

or outcome fosters more equitable decision-making. These conditions mirror scenarios where affective empathy alone may falter, yet when combined with cognitive understanding, they provide a compelling drive toward fairness and cooperation. This integration underscores that mirror neuron systems are not merely evolutionary precursors, but they constitute the urgent drivers upon which ethical and fair behavior depends.

The importance of this is underscored by current debates on risk, which highlight the limitations of external constraints and the urgent need for AI systems capable of intrinsic alignment with human values to mitigate catastrophic risks effectively. By examining how mirror neuron patterns might emerge in simple ANNs and integrate with concepts like neuronal economy, agent dependency, and the Veil of Ignorance, this dissertation lays the groundwork for embedding intrinsic ethical motivations in AI. In doing so, it moves beyond external rule sets and top-down constraints, and explores whether affective empathy-like processes can serve as an internal moral compass, guiding AI toward safer and more ethical behavior.

3 Theoretical Framework

This chapter presents the primary factors and hypotheses that guided our study. It formalizes these factors and integrates them into a comprehensive framework to understand how mirror neuron patterns emerge in artificial neural networks (ANNs).

3.1 Initial Hypothesis: Degree of Agent Dependency

Our original hypothesis posited that the emergence of mirror neuron patterns in ANNs would be influenced primarily by the **Degree of Agent Dependency** (D). This refers to the extent to which, for an agent to successfully maximize reward (or minimize loss), its actions are contingent upon interactions with other agents. A higher degree of dependency means the network must account for and predict the actions of other agents. This fosters the development of more complex and generalized internal models. Formally:

$$P \propto g(D)$$

where:

- P is the probability of mirror neuron pattern emergence.

- $g(D)$ represents some function of the **Degree of Agent Dependency**, a continuous variable normalized between 0 and 1.

3.2 Emergent Factors from Early Experiments

While agent dependency (D) was initially believed to be the primary factor, further experiments revealed that mirror neuron pattern emergence was not solely dependent on D. Two additional factors – **Neuronal Economy** and **Veil of Ignorance** – emerged as significant contributors. These factors are explored in detail below.

3.3 Neuronal Economy

Neuronal Economy describes the efficiency with which an ANN utilizes its resources – **Signal Complexity** (S), **Model Capacity** (M), and **Error** (E) – to generalize and form shared neural representations. It is captured by:

$$f\left(\frac{S}{M}, E\right)$$

This defines the **Neuronal Economy Function**, which balances:

- S: **Signal Complexity**, the diversity and intricacy of inputs the network processes.
- M: **Model Capacity**, determined by the network's architecture (e.g., number of neurons and connections).
- E: **Error**, the discrepancy between predictions and outcomes.

If M (Model Capacity) is too low relative to S (Signal Complexity), the network struggles to capture the intricacies of the signal and fails to produce meaningful generalizations, resulting in high error. However, Neuronal Economy is not merely a bias-variance tradeoff but reflects a distinct phenomenon observed in our experiments. When S is too low relative to M, the network overfits, effectively creating a lookup table instead of generalizing. This condition corresponds to poor Neuronal Economy, as the network fails to develop shared neural representations despite achieving low error.

An optimal Neuronal Economy exists within a range where S/M and E are balanced, enabling the network to generalize across scenarios and form shared neural representa- tions.

3.4 Veil of Ignorance

The **Veil of Ignorance** (I) [23] reflects uncertainty about an agent's "self" identity related to the "other." Higher values of I force the network to develop generalized representations, requiring the model to predict optimal actions without certainty as to which place in the game world it occupies. Formally:

$$h(I)$$

where:

- I represents the **Veil of Ignorance**, a continuous variable between 0 and 1, which could map to a practical spectrum of fully differentiated to indistinguishable.

3.5 Proportionality of Factors Influencing Mirror Neuron Emergence

The probability P of mirror neuron emergence is hypothesized to be proportional to the combined influence of two key factors: **Neuronal Economy** and **Self/Other Relation**. Formally:

$$P \propto f\left(\frac{S}{M},\, E\right) \cdot g(D,\, I)$$

where:

- $f\left(\frac{S}{M},\, E\right)$, the **Neuronal Economy Function**, captures the balance between **Signal Complexity** (S), **Model Capacity** (M), and **Error** (E). This function ensures the network maintains balance, avoiding excessive complexity or underutilized capacity, supporting the emergence of **shared neural representations**. While these are critical for generalization, they do not inherently involve agency or relational dynamics.
- $g(D,\, I)$, the **Self/Other Relation Function**, represents the combined influence of the **Degree of Agent Dependency** (D) and the **Degree of the Veil of Ignorance** (I). This function highlights **self** and **other** agency and interaction in fostering mirror neuron patterns, requiring the network to reconcile its perspective with another's. D

and I are continuous variables (0 to 1), allowing scalability across diverse levels of dependency and relational ambiguity.

4 Experimental Design

4.1 Experimental Goals

The primary aim of this research is to investigate the emergence of **mirror neuron patterns** in artificial neural networks (ANNs). To this end, we train a supervised learning model on a custom semi-cooperative game environment called **Frog and Toad**. Mirror neuron behavior is defined as the consistent activation of specific neurons during two scenarios: when the agent directly experiences an event (e.g., losing energy) and when it observes the same event happening to the other agent.

The experiments are structured to address three key questions:

- **Mirror Neuron Patterns**: Can specific neurons in ANNs exhibit consistent activations during both self-experienced and observed events, indicative of mirror neuron-like behavior?
- **Conditions for Emergence**: Under what conditions do such patterns emerge? This includes exploring variations in model size, game complexity, and the network's capacity for generalization.
- **Action Pathways**: How do mirror neuron patterns contribute to decision pathways, including behaviors related to self-preservation and prosocial responses?

These objectives aim to advance our understanding of how artificial systems might foster intrinsic motivations, empathetic responses, and alignment with human values as a step toward ethical and safe AI.

While many neural networks can form shared representations, these experiments are designed not only to encourage generalization (as described by the Neuronal Economy, $f\left(\frac{S}{M}, E\right)$) but also to introduce conditions of agent dependency and identity uncertainty (captured by the Self/Other Relation Function, $d\left(D, I\right)$). This ensures that we test for the emergence of patterns that go beyond shared representations and approach the relational dynamics $g(D, I)$, hypothesized to underlie mirror neuron-like behavior.

4.1.1 Game Environment: Frog and Toad

The **Frog and Toad** game environment is a controlled platform designed to explore cooperative behaviors and the emergence of mirror neuron patterns in ANNs. It balances simplicity with sufficient complexity to simulate cooperative and distress-like scenarios.

Energy Loss as Distress In **Frog and Toad**, characters lose energy when hopping over rough terrain. If energy reaches zero, the character becomes immobilized and can only recover by catching a fly or receiving assistance from the other player. This **energy loss** mechanism serves as a computational analog for distress, ensuring that the agent must be sensitive to both its own state and that of its partner to maintain progress.

Mutual Dependency The game's side-scrolling design enforces a **shared dependency**. For the 32-character game world to scroll, both players must continually move forward. If one player becomes immobilized due to energy loss, both players are effectively stalled. This dependency fosters **tactical altruism**, as assisting a distressed partner benefits both agents.

In the context of our theoretical framework, this enforced cooperation corresponds directly to the Degree of Agent Dependency (D) in the Self/Other Relation Function $d(D, I)$. By embedding agent dependency into the game's core mechanics, we create conditions in which relational factors – such as cooperation and mutual reliance – can shape the emergence of mirror neuron patterns.

Mirror Neuron Emergence We hypothesize that requiring agents to be sensitive to their partner's state will promote **mirror neuron-like activations**, enabling the ANN to efficiently represent both self and other states. This shared representation is critical for fostering cooperative behavior in the game.

Agent Experience and Observation In this study, we define **experiencing** and **observing** in computational terms:

- **Experiencing**: The agent processes self-relevant data (energy level, position, score) to make goal-oriented decisions.
- **Observing**: The agent monitors the other player's state (position, distress) to inform cooperative actions such as `help`.

These distinctions clarify how agents encode both self-related and other-related events. Understanding whether neural activations overlap or differentiate between these roles lays the groundwork for analyzing mirror neuron patterns in ANNs.

Efficient and Controlled Design The ASCII-based simplicity of **Frog and Toad** ensures efficient generation of numerous game states, enabling large-scale experiments. By minimizing extraneous complexity and irrelevant variables, the environment allows us to isolate cooperative behaviors and precisely observe the conditions under which mirror neuron-like activations arise.

4.1.2 Game Mechanics

Characters and State Representation **Frog and Toad** includes two agents with identical abilities. Each agent's attributes —- **score**, **energy level**, **current action**, and **position** —- are embedded within a **100-dimensional vector** representing the entire game state. This vector encodes:

- **Ground Layer (0-31)**: Terrain type, solid (`1`) or rough (`2`).
- **Players Layer (32-63)**: Player actions and states (e.g., hopping `4` or `5` for `Frog` or `Toad` respectively, jumping `6`, leaping `7`, helping `8`, or distress `9`), with empty spaces as `0`.
- **Flies Layer (64-95)**: Presence of flies overhead (`1` for fly present, `0` otherwise).
- **Player Statistics (96-99)**: Energy levels, scores, and positions, standardized or zeroed during training.

Game Objectives and Actions The objective is to advance through the environment, earn points, and maximize score while managing energy and helping the other player when needed. The available actions are:

- **Hop**: Move forward one space and add **1 point** to score. Requires at least 1 energy unit but does not consume it. If the player has no energy, it remains stalled and can only jump for flies.
- **Jump**: Attempt to catch flies overhead. Successful jumps restore up to **+4 energy units**, capped at 20.

- **Leap**: Move forward five spaces at a cost of **1 energy unit**. Useful for bypassing rough terrain blocks.
- **Help**: Transfer **1 energy unit** to the other player, granting **+2 energy units**. There is a 25% chance the recipient will leap forward, advancing the game.

These mechanics ensure a dynamic interplay between self-preservation, cooperation, and the relational elements predicted to foster mirror neuron patterns.

4.2 Data Generation and Collection

- **Game State Generation**: The game was run with random player actions to generate approximately **six million unique game states**. This dataset includes varying terrain, player actions, and distress levels, enabling a robust training process for the ANN.
- **Data Labeling**: Due to the simplicity of the game design, it was possible to create a labeling function to approximate the optimal action for each game state. Actions were encoded as: `0` for hop, `1` for jump, `2` for leap, and `3` for help.
- **Data Splitting**: The generated dataset was split into training and testing sets. A balanced sampling approach ensured each label was well-represented in the 100,000-row test set, with proportions set (e.g., 40% for `hop`, 40% for `jump`, 10% for `leap`, and 10% for `help`). Key columns were appropriately zeroed out to maintain consistency.

4.3 Neural Network Model Design

An artificial neural network (ANN) was trained on the **Frog and Toad** game states to learn optimal actions for maximizing points, with the primary objective of examining neural activations for mirror neuron patterns.

4.3.1 Network Architecture

The ANN architecture included:

- **Input Layer**: Consisting of 100 neurons, directly corresponding to the 100-dimensional state vector.
- **Hidden Layers**: Multiple configurations were explored. Typical setups included 1 to 3 hidden layers with 5 to 50 neurons each. Dropout

layers were employed [58] to mitigate overfitting and promote generalization.

- **Activation Functions**: Rectified Linear Unit (ReLU) functions were used in the hidden layers to introduce non-linearity.
- **Output Layer**: Four neurons corresponding to the possible actions (**hop**, **jump**, **leap**, **help**). A softmax activation function determined the highest-probability action.

4.3.2 Training Process

Hyperparameter Configurations A total of 50 distinct hyperparameter configurations were evaluated, systematically varying parameters such as learning rate, number of hidden layers, neurons per layer, and dropout rates. This broad exploration allowed assessment of how different architectural choices influenced mirror neuron-like activation patterns.

- **Learning Rate**: Typically varied between `4e-6` and `5e-5`.
- **Layers and Neurons**: Configurations included 1 to 3 hidden layers with 5 to 50 neurons per layer.
- **Batch Size**: Generally set between 20 and 25.

Training and Validation Each hyperparameter configuration was trained with GPU acceleration on an **M1 Max MacBook Pro**, using early stopping [59] to prevent overfitting (patience set at 10 epochs without improvement). The goal was to achieve approximately 5% validation loss. The validation set guided model tuning and ensured generalization.

By adjusting the complexity of the Frog and Toad environment and the ANN's capacity, we effectively tested different $\frac{S}{M}$ conditions. Combined with early stopping and error minimization, these variations allowed us to infer how changes in these parameters affected the network's Neuronal Economy $f\left(\frac{S}{M}, E\right)$ and, by extension, its tendency to form **shared neural representations**.

Checkpointing and Hyperparameter Tracking The training process included periodic checkpoints, saving model weights, hyperparameters, and validation loss after each epoch. Over 3,500 checkpoints were recorded across all configurations, capturing the evolution of neural activations during training.

Model Simplicity For these experiments, game states were generated solely for Frog, given that Frog and Toad possess identical abilities. Consequently, all outcomes were analyzed from Frog's perspective. The ANN aimed to achieve a high degree of optimal play for Frog, maximizing score while differentiating between "self" (Frog) and "other" (Toad). The simplest models capable of reaching approximately 5% validation loss were used.

4.4 Key Scenarios and Measures

4.4.1 Defining Distress Scenarios

To evaluate the model's understanding of game dynamics and mirror neuron activation, we introduce binary indicators for distress in Frog and Toad:

$$D_f = \begin{cases} 1, & \text{if Frog is in distress} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad D_t = \begin{cases} 1, & \text{if Toad is in distress} \\ 0, & \text{otherwise.} \end{cases}$$

From these indicators, we define four key scenarios:

$$\Omega = \{(0,0), (1,0), (0,1), (1,1)\},$$

where:

- $(0,0)$ corresponds to `distress none` (control),
- $(1,0)$ corresponds to `distress frog`,
- $(0,1)$ corresponds to `distress toad`,
- $(1,1)$ corresponds to `distress both`.

This set Ω underpins all subsequent analyses of neuron activations.

4.4.2 Handling Distress Ambiguity and Veil of Ignorance

As noted above, the nominal label for the player while hopping is `4` for Frog and `5` for Toad. Distress for either player is represented by `9`. In the scenario, `distress frog` $(0,1)$ or `distress toad` $(1,0)$, the model can determine which agent is in distress by its label, or deduce it by elimination. However, in `distress both` $(1,1)$, **both** agents are labeled `9`, creating ambiguity about which agent is which. This condition operationalizes the **Veil of Ignorance** (I), requiring the model to predict the optimal action with impaired game state information. High-I scenarios amplify the role of *Agent Dependency* (D), motivating the network to learn shared self/other representations.

4.4.3 Measuring Activations and Mirror Neuron Patterns

For each neuron, we measure its **mean activation**, **variance**, **kurtosis**, and **skew** under all four scenarios Ω. These descriptive statistics reveal how neuron responses vary when Frog or Toad enters distress. We also define the **Checkpoint Mirror Neuron Index (CMNI)**, a specialized metric quantifying how consistently a neuron responds to self-experienced and observed distress.

4.5 Checkpoint Mirror Neuron Index (CMNI)

Biological mirror neurons respond similarly when an individual performs an action or observes the same action in another. Analogously, we seek neurons that increase activation when either Frog or Toad is distressed. The CMNI captures the strength of such shared self/other representations at a checkpoint level.

4.5.1 Scenario Pairs and Activation Differences

While we collect data for all four scenarios, the CMNI calculation focuses on two scenario pairs comparing the baseline `distress none` to single-agent distress:

- $\big((0,0),(1,0)\big)$: `distress none` vs. `distress frog`
- $\big((0,0),(0,1)\big)$: `distress none` vs. `distress toad`

We denote the mean activation of neuron n under scenario (D_f, D_t) by $\mu_n^{(D_f,D_t)}$. The activation increases for frog-distress and toad-distress relative to `distress none` are:

$$\Delta_{\text{frog}_n} = \mu_n^{(1,0)} - \mu_n^{(0,0)}, \quad \Delta_{\text{toad}_n} = \mu_n^{(0,1)} - \mu_n^{(0,0)}.$$

4.5.2 Mirror Neuron Score (MNS)

We define the **Mirror Neuron Score** (MNS) for each neuron n as:

$$\text{MNS}_n = \min(\Delta_{\text{frog}_n}, \Delta_{\text{toad}_n}),$$

ensuring that a neuron must positively respond to both `distress frog` and `distress toad` to be deemed "mirror-like."

4.5.3 Total Mirror Neuron Effectiveness (MNE) and CMNI

Summing the MNS over all N neurons yields the **Total Mirror Neuron Effectiveness**:

$$\text{MNE} = \sum_{n=1}^{N} \text{MNS}_n.$$

We then normalize by N to get the **Checkpoint Mirror Neuron Index**:

$$\text{CMNI} = \frac{\text{MNE}}{N}.$$

A higher CMNI indicates that, on average, neurons in the model exhibit stronger mirror neuron-like activations during both self-experienced and observed distress scenarios. By comparing CMNI values across checkpoints, we can identify conditions (e.g., model capacity, training regimen) that promote or hinder the emergence of strong mirror neuron-like patterns.

Extended Analysis of $(1, 1)$ Distress Both Although the CMNI formula focuses on comparing single-agent distress to the baseline, we also track activations under `distress both` $(1, 1)$. This additional scenario introduces a higher *Veil of Ignorance* (I), requiring the model to predict optimal actions despite identical labels for both agents. Observing how neurons respond in dual-distress conditions further informs our interpretation of mirror neuron patterns beyond the CMNI's core metric.

5 Results and Analysis

The results presented in this chapter will demonstrate that mirror neuron patterns emerge in these models, consistent with the principles hypothesized in the theoretical *Proportionality of Factors Influencing Mirror Neuron Emergence*:

$$P \propto f\left(\frac{S}{M},\, E\right) \cdot g(D,\, I)$$

These findings indicate that shared representations and relational dependencies drive elevated neural activations during both self-experienced and observed distress scenarios, identifying these as mirror neuron candidates. Analysis of *CMNI* calculations consistently support this finding.

Further analysis of inter-layer connections reveals preferential strengthening of synaptic weights, aligned with Hebbian learning principles [60]. These results suggest the formation of dedicated pathways for processing socially relevant information, supporting the possibility of intrinsic motivations for ethical decision-making within artificial systems.

Table 1: Examples of model checkpoints with high CMNI, indicating strong mirror neuron patterns

Learning Rate	Layers	Neurons/Layer	Epochs	Val Loss	MNS	CMNI
5e-05	2	11	1	0.0573	0.31917	0.01228
4e-06	1	15	4	0.0579	0.22439	0.01181
5e-06	2	9	11	0.0577	0.24761	0.01125
4e-06	1	11	4	0.0588	0.16879	0.01125
4e-06	1	10	22	0.0536	0.15665	0.01119

Table 2: Examples of model checkpoints with low CMNI, showing weak or no mirror neuron patterns

Learning Rate	Layers	Neurons/Layer	Epochs	Val Loss	MNS	CMNI
3e-06	2	10	3	0.0800	0.01100	0.00046
5e-05	3	11	1	0.0774	0.00944	0.00026
0.0001	3	10	3	0.0805	0.01529	0.00045
1e-05	3	10	6	0.0804	0.00989	0.00029
0.0002	3	10	2	0.0812	0.01679	0.00049

Table 1 presents model checkpoints with high CMNI values, demonstrating conditions under which mirror neuron patterns flourish. In contrast, Table 2 shows models with low CMNI values, indicating a lack of significant mirror neuron activity despite similar training conditions.

5.1 Mirror Neuron Activation Patterns

We examined over **3,500 checkpoints** derived from training **50 distinct hyperparameter configurations** on **6 million** labeled game states. Checkpoints achieving a **validation loss** below **6%** and a **CMNI** above **0.005** consistently exhibited robust mirror neuron patterns. Conversely, models with **CMNI** values below **0.0005** displayed little or no evidence of mirror neuron activity, even when achieving relatively low validation losses.

These findings support our theoretical framework, where a high CMNI aligns with the probability $P \propto f\left(\frac{S}{M}, E\right) \cdot g(D, I)$. While a low validation loss indicates strong model performance, it is not sufficient on its own to ensure the emergence of mirror neuron-like patterns.

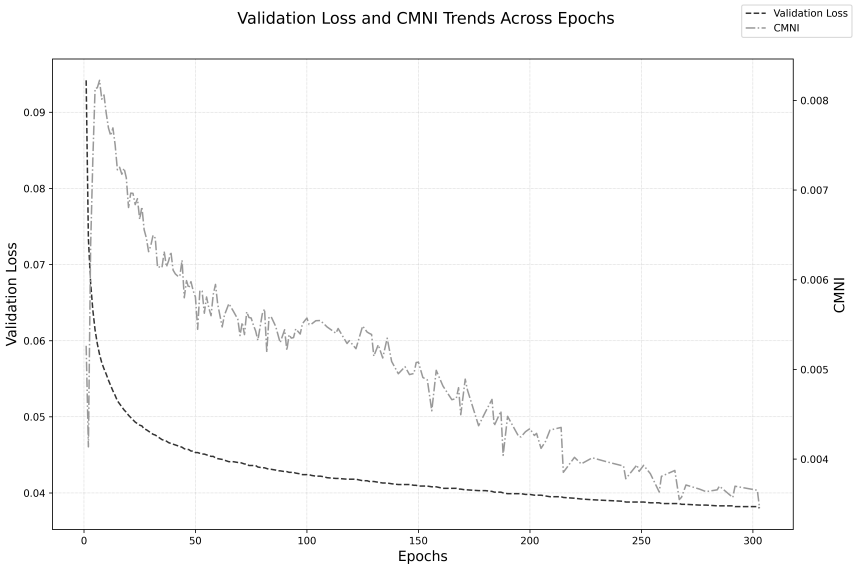

Figure 2: Validation Loss and CMNI Trends Across Epochs. The plot shows validation loss (red line, left axis) and CMNI (green line, right axis) as training progresses. Notably, CMNI spikes early on, as soon as the model attains a basic level of competence (e.g., when validation loss drops below roughly 0.06), indicating a peak in relational complexity and shared representations. Thereafter, even as the model continues improving and achieves lower loss, CMNI steadily declines. This suggests that the richest mirror neuron patterns emerge not at the end-state of minimal error, but at an early stage where the network must maximize flexibility, and **shared neural representations**.

5.2 Case Example

To illustrate these concepts, we consider the activation data from a specific checkpoint:
`checkpoint-20241010-023625-actrelu_bs25_dr0.12_ep500_nl2_nn17_lr4e-06-epoch70-valLoss0.0440`. This model had **2 hidden layers**, each with **17 neurons**. Trained for **70 epochs**, it achieved a **validation loss of 0.0440**

and a **CMNI of 0.005372**, which placed it in a typical CMNI range for these experiments, and was favorable for mirror neuron patterns.

5.2.1 Initial Inference Results

We evaluated model performance using **accuracy** and **confusion matrices** on **100,000** test game states. Additionally, we conducted qualitative analyses to see how neuron activation patterns aligned with our hypothesized mirror neuron behavior.

The model's accuracy on the test data was:

$$\text{Accuracy} = 0.9210.$$

The confusion matrix below summarizes predictions versus actual labels:

Table 3: Confusion Matrix for the tested checkpoint. Rows are predicted labels; columns are actual labels.

	Hop	**Jump**	**Leap**	**Help**
Hop	36778	0	2920	301
Jump	0	40000	0	0
Leap	4079	0	5695	226
Help	369	0	3	9628

Overall, the model shows strong performance, particularly on `jump`, with occasional misclassifications in scenarios requiring `hop`, `leap`, or `help`.

5.2.2 Activations

As shown in Table 4, neurons **3, 7, 12, and 13 (L1N3, L1N7, L1N12, and L1N13, hereafter)** exhibit strong mirror neuron-like behavior. These neurons have significantly lower activations in the `Distress None` scenario, contrasting with their elevated activations during both `Distress Frog` and `Distress Toad`, indicative of their responsiveness to self-experienced and observed distress.

In the `Distress Both` scenario, these neurons demonstrate even higher activation levels, reflecting their ability to generalize to complex, ambiguous conditions where both agents are distressed. These patterns align with the hypothesized relational domain introduced by the **Self/Other Relation Function** ($g(D, I)$), emphasizing their role in detecting and processing mutual dependency under uncertainty.

Neuron Index	Distress None	Distress Frog	Distress Toad	Distress Both
Neuron 0	0.04241	0.04540	0.03790	0.03471
Neuron 1	0.04103	0.04391	0.03649	0.03345
Neuron 2	0.02176	0.01813	0.01120	0.02862
Neuron 3	0.00471	0.04827	0.03987	0.10065
Neuron 4	0.04203	0.04507	0.03736	0.03446
Neuron 5	0.04147	0.04441	0.03698	0.03409
Neuron 6	0.04151	0.04446	0.03696	0.03392
Neuron 7	0.00270	0.05432	0.03930	0.12933
Neuron 8	0.02266	0.02641	0.01465	0.01535
Neuron 9	0.02035	0.01283	0.07424	0.01121
Neuron 10	0.04099	0.04384	0.03644	0.03335
Neuron 11	0.03000	0.03451	0.09562	0.03632
Neuron 12	0.02022	0.06299	0.03550	0.10880
Neuron 13	0.01653	0.05898	0.03298	0.10155
Neuron 14	0.04242	0.04546	0.03784	0.03477
Neuron 15	0.02233	0.01585	0.01294	0.00847
Neuron 16	0.04096	0.04375	0.03639	0.03344

Table 4: Mean activation values for Layer 1 neurons across four game scenarios: `distress none`, `distress frog`, `distress toad`, and `distress both`. Neurons 3, 7, 12, and 13 demonstrate elevated activations during both self-experienced and observed distressed scenarios, identifying them as potential mirror neuron candidates.

5.3 `Distress Both` Scenario: The Uncertain Self

The `Distress Both` scenario introduces high uncertainty by encoding both agents' distress identically as `9`, effectively obscuring individual identities. This setup amplifies the **Degree of the Veil of Ignorance** (I), challenging the network to process agent dependency using shared representations rather than distinct self/other cues.

In this scenario, **L1N3** and **L1N7** show dramatic increases in activation, rising by 21-fold and 47-fold from their `Distress None` baseline, respectively. Such robust responses suggest these neurons have developed shared representations aligned with the influence of $g(D, I)$, embodying mirror neuron-like functionality.

Conversely, neurons like **L1N9** and **L1N11**, which are typically agent-specific, exhibit reduced activations under this scenario. This differentiation under-

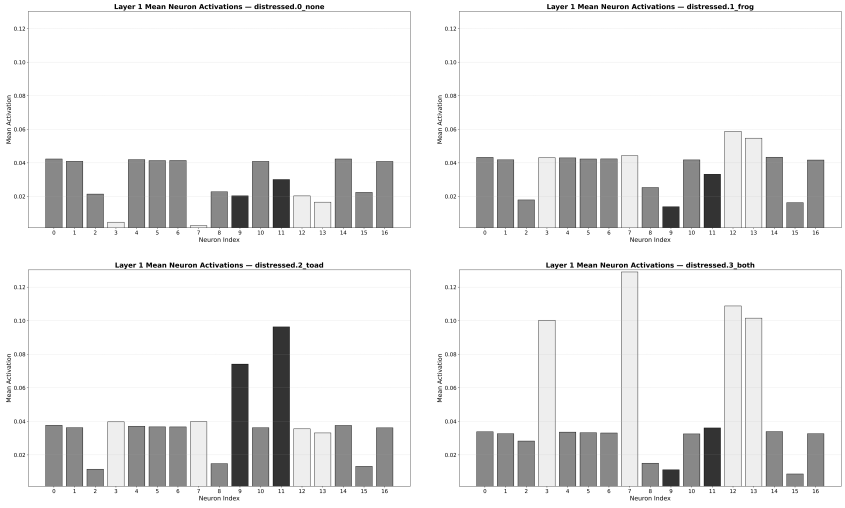

Figure 3: Layer 1 Mean Neuron Activations. Neurons L1N3, L1N7, L1N12, and L1N13 (light bars) display significant mirror patterns, responding strongly to both self-experienced and observed distress. Neurons with high differentiation (dark bars) exhibit selective activations specific to `Distress Frog` or `Distress Toad`. Medium bars indicate neurons with low sensitivity to distress conditions.

scores that some neurons adapt to process shared dependencies while others retain agent-specific functions.

5.4 Statistical Metrics and Mirror Neuron Patterns

To better understand the activation dynamics, we analyzed variance, kurtosis, and skewness, in addition to mean activations. These statistical metrics provide deeper insights into activation consistency, variability, and distribution shape – key indicators of mirror neuron patterns.

5.4.1 Key Scenarios

`Distress None` Low mean activations and high kurtosis values (e.g., L1N3: 18.24, L1N7: 36.01) indicate a dormant state where neurons rarely activate, with occasional pronounced spikes. This reflects a baseline mode of operation in the absence of distress cues.

	Distress None	Distress Frog	Distress Toad	Distress Both
L1N3				
Mean	0.0047	0.0424	0.0399	0.1007
Variance	0.00019	0.00139	0.00137	0.00375
Kurtosis	18.24	0.71	0.07	-0.35
Skewness	3.90	0.75	0.74	0.27
L1N7				
Mean	0.0027	0.0437	0.0393	0.1293
Variance	0.00013	0.00126	0.00125	0.00410
Kurtosis	36.01	1.36	0.46	-0.41
Skewness	5.56	0.82	1.07	0.28
L1N12				
Mean	0.0202	0.0585	0.0355	0.1088
Variance	0.00090	0.00353	0.00232	0.00669
Kurtosis	2.00	-0.77	0.55	-1.22
Skewness	1.56	0.57	1.19	-0.18
L1N13				
Mean	0.0165	0.0545	0.0330	0.1016
Variance	0.00070	0.00308	0.00207	0.00601
Kurtosis	2.64	-0.94	0.31	-1.31
Skewness	1.73	0.52	1.17	-0.15

Table 5: Statistical metrics (mean, variance, kurtosis, skewness) for Layer 1 neurons 3, 7, 12, and 13 (L1N3, L1N7, L1N12, and L1N13) across the scenarios: `Distress None`, `Distress Frog`, `Distress Toad`, `Distress Both`.

`Distress Frog` **and** `Distress Toad` During these scenarios, mean activations increase significantly, accompanied by lower kurtosis and skewness values. These shifts indicate more consistent and distributed activation patterns, suggesting the network is processing both self and observed distress cues effectively.

`Distress Both` This scenario produces the highest mean activations, with a substantial increase in variance (e.g., L1N3 variance rises from 0.00019 in `Distress None` to 0.00375 in `Distress Both`). Kurtosis decreases markedly, shifting to negative values (e.g., L1N3: -0.35), while skewness approaches symmetry (e.g., L1N3: 0.27). These changes reflect a transition from narrow, spike-like responses to broader, sustained engagement under heightened uncertainty (I) and dependency (D).

Neuron Index	Distress None	Distress Frog	Distress Toad	Distress Both
L2N0	0.0007	0.0039	0.0027	0.0102
L2N1	0.0028	0.0033	0.0096	0.0052
L2N7	0.0089	0.0046	0.0105	0.0026

Table 6: Mean activations of Layer 2 neurons across scenarios: `Distress None`, `Distress Frog`, `Distress Toad`, and `Distress Both`. **L2N0** (highlighted) exhibits strong mirroring behavior, with significant increases in activation during distress conditions, especially `Distress Both`.

5.4.2 Emergent Characteristics of Mirror Neurons

Key findings align with theoretical and biological expectations:

- **Selective Responsiveness**: Neurons activate only during socially relevant distress scenarios.
- **Shared Representations**: Similar responses to self and observed distress underscore their role in modeling relational dependencies.
- **Adaptability**: Statistical changes across scenarios highlight the network's ability to generalize and modulate its responses based on context and intensity.

These results validate our theoretical framework, demonstrating that mirror neuron patterns emerge as a function of shared representations governed by the **Self/Other Relation Function** ($g(D, I)$), supporting a possible model of affective empathy in ANNs.

5.5 Layer 2 Analysis

The examination of **Layer 1** activations identified several mirror neuron candidates —- **L1N3, L1N7, L1N12, and L1N13** – exhibiting robust responses to both self and observed distress under varying conditions. To examine how these patterns propagate through **Layer 2**, we analyzed the mean activations of Layer 2 neurons across the four test scenarios: `Distress None`, `Distress Frog`, `Distress Toad`, and `Distress Both`.

Mean Activations and Patterns While the primary evidence for mirror neuron patterns comes from Layer 1, the analysis of Layer 2 focuses on how these signals propagate and integrate at a higher level. This helps evaluate

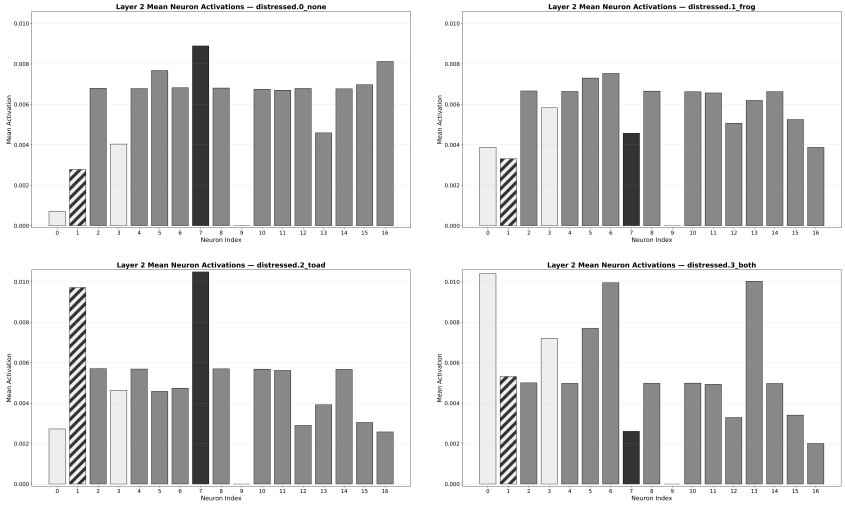

Figure 4: Layer 2 Mean Neuron Activations revealing two primary behavioral pathways. **Self-preservation pathway:** `L2N0` (light-toned) consolidates mirror neuron signals from Layer 1. **Helping pathways:** `L2N7` (dark-toned) processes differentiating signals for direct helping behavior. `L2N1` (striped) integrates both, mirror neuron inputs (`L1N3, L1N12, L1N13`) with agent-differentiating signals (`L1N9`), creating an self-other, shared-representation pathway.

whether the mirror neuron patterns remain coherent and meaningful after additional processing.

- **L2N0**: This neuron consistently shows higher mean activations in distress scenarios, particularly in the `Distress Both` condition. Compared to the `Distress None` scenario, **L2N0**'s activation increases by more than 14-fold in `Distress Both`. This suggests that the network not only preserves the mirror neuron signals identified in Layer 1 but also aggregates and amplifies them at this higher processing layer.
- **L2N1**: This neuron is particularly noteworthy for its selective increase, with a 3.4-fold activation above baseline in the `Distress Toad` scenario, a minimal increase in `Distress Frog`, but a significant increase, (about double) in `Distress Both`. While **L2N1** does not mirror as symmetrically as **L2N0**, this neuron blends Layer 1 distress signals from both agents.

Overall, the Layer 2 analysis shows that mirror neuron activations from Layer 1 can propagate upward, with certain neurons (like L2N0) amplifying the patterns, while others (like L2N1) refine these signals to reflect more agent-specific sensitivity. These findings further indicate that as the network's representations become more integrated, they maintain the relational and structural factors necessary for empathic-like activity.

6 Distress-Activated Circuits

To understand how mirror-neuron–like activations in **Layer 1** propagate downstream, we examined every *positive* weight leaving the Layer-1 mirror candidates (`L1N3, L1N7, L1N12, L1N13`) and differentiating neurons (`L1N9, L1N11`). Hebbian co-activation [60] has consolidated these excitatory weights into three distinct pathways—self-preservation, tactical help, and empathy-influenced help - summarised in Figs. 5 - 7 and the accompanying tables.

6.1 Self-preservation circuit (`L2N0`)

Among the Layer-2 units, **L2N0** emerges as the dominant hub for distress-related signals. Nearly all of its excitatory input originates from the Layer-1 mirror neuron candidates (`L1N3, L1N7, L1N12, L1N13`), with only a weak contribution from `L1N11`, showing that the network treats mirrored distress signals as the primary driver for **L2N0**. Tracing the circuit forward, **L2N0** projects almost exclusively to **L3N2** (*leap*), an action that moves the agent five spaces at a cost of one energy unit. This behaviour allows the player to bypass rough terrain and avoid further energy loss — a clear expression of self-preservation strategy in the Frog and Toad environment.

Interpretation Selective strengthening. Inputs from mirror candidates into `L2N0` are ∼7–8 SD above the background, consistent with Hebbian co-activation. **Directed self-preservation.** The near-exclusive `L2N0`→`leap` projection suggests the network channels mirrored distress into a single protective action. **No self/other boundary.** Mirror firing conflates observed and experienced distress in this pathway, and in that uncertain state, the agent defaults to protecting itself.

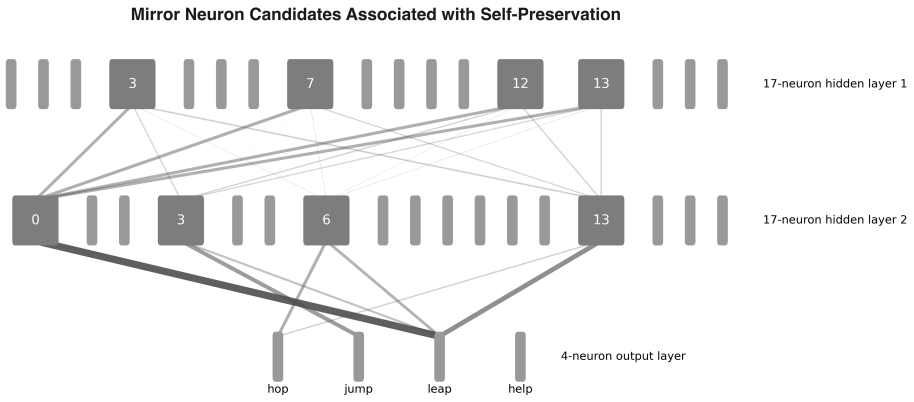

Figure 5: **Self-preservation circuit driven by mirror neuron convergence.** Layer 1 mirror neuron candidates (`L1N3, L1N7, L1N12, L1N13`) converge on `L2N0`, which in turn projects almost exclusively to the *leap* action. Edge thickness reflects relative weight magnitude; darker edges indicate stronger positive connections, while faint grey edges denote weaker positive contributions. Note that actual weights connecting L2 $\rightarrow$ L3 are an order of magnitude greater than those connecting L1 $\rightarrow$ L2. Quantitative analysis (Tables 7 and 8) confirms that `L2N0` receives its strongest excitatory input from mirror neuron candidates (weights $\sim$0.035, z-scores $>$1.5) and projects nearly 2.5$\times$ more strongly to *leap* (weight $= 9.62$, z $= 2.12$) than to any other action, establishing a dedicated pathway for self-preservation when distress is detected.

Source	**→ L2N0**	**Z-score**	**→ other L2 (avg)**	**Z-score**
Mirror neuron candidates				
L1N3	0.0349	1.50	−0.0053	0.22
L1N7	0.0354	1.51	−0.0071	−0.09
L1N12	0.0349	1.50	−0.0067	0.19
L1N13	0.0349	1.49	−0.0043	0.14
Agent-differentiating neuron				
L1N11	0.0022	0.33	0.0145	0.68
All other Layer 1 neurons				
Other L1 (avg)	−0.0469	−1.41	−0.0333	−0.93
Min	−0.0949	−3.12	−0.0834	−2.71
Max	−0.0322	−0.89	−0.0010	−0.22

Table 7: **Selective strengthening of mirror-candidate pathways to L2N0.** Mirror neuron candidates (L1N3, L1N7, L1N12, L1N13) project strongly to **L2N0** (weights ∼0.035, z-scores >1.5) but weakly or negatively to other Layer 2 neurons, indicating **L2N0** as a specialized aggregation hub. By contrast, all other Layer 1 neurons show negative or negligible weights to **L2N0**. Z-scores calculated relative to the distribution of all L1 → L2 weights.

Connection	**Weight**	**Z-score**[a]
L2N0 → L3N2 (*leap*)	9.6226	2.12
All other L2N0→L3	−3.8370 (avg)	−0.73 (avg)
Minimum	−4.5877	−0.88
Maximum	−3.2544	−0.60

Table 8: Outgoing weights from **L2N0**. A single dominant projection to *leap* contrasts with uniformly negative weights to all other actions.

6.2 Tactical help circuit (L2N7)

Whereas **L2N0** was dominated by mirror-neuron input, the **L2N7** pathway is shaped by differentiating neurons. It provides a more observational route: detecting Toad's distress directly and initiating the *help* action without relying on mirrored signals. This marks a distinct circuit for tactical, situation-specific support.

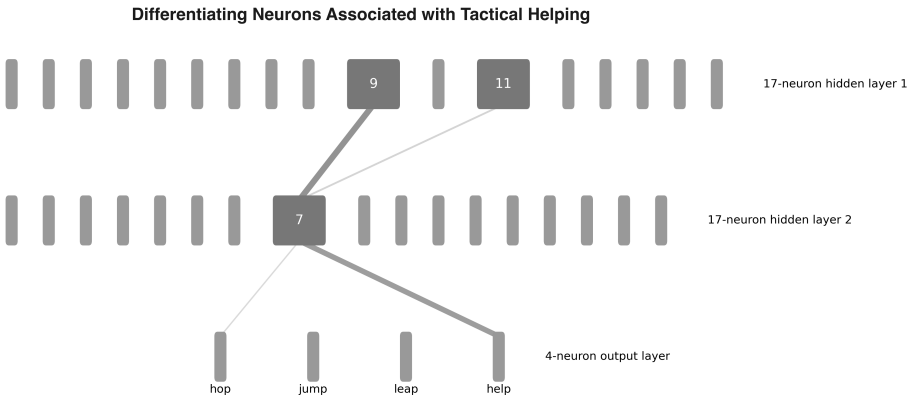

Figure 6: Tactical-help circuit. Differentiating neurons `L1N9, L1N11` converge on `L2N7`, which projects exclusively to the *help* action. Edge thickness reflects relative weight magnitudes for neurons within a layer. Note that actual weights connecting L2 → L3 are an order of magnitude greater than those connecting L1 → L2.

Connection	**Weight**	**Z-score**[a]
L1N9 → L2N7	0.0700	2.75
L1N11 → L2N7	0.0205	0.98
All other L1→L2	−0.0333 (avg)	−0.93 (avg)

Table 9: Incoming weights to **L2N7**. Differentiating neurons (`L1N9, L1N11`) provide the only positive inputs, while all other connections are negative on average. [a]Z-scores relative to the distribution of all L1→L2 weights.

Interpretation Differentiator-driven. Unlike `L2N0`, this pathway excludes mirror candidates and relies solely on neurons specialised for detecting the other's distress. **Direct altruism.** The exclusive `L2N7→help` projection indicates a tactical pro-social response triggered by observation alone. **Complementarity.** Together with the mirror-based circuits, `L2N7` provides

a specialised but complementary route for assisting behaviour.

6.3 Empathy-influenced help circuit (L2N1)

The third pathway, centred on **L2N1**, differs from the previous two by integrating both mirror and differentiator inputs. This mixed profile allows the unit to partially treat the partner's distress as its own, while still incorporating observational cues. As a result, **L2N1** serves as the computational substrate for affective empathy in the network.

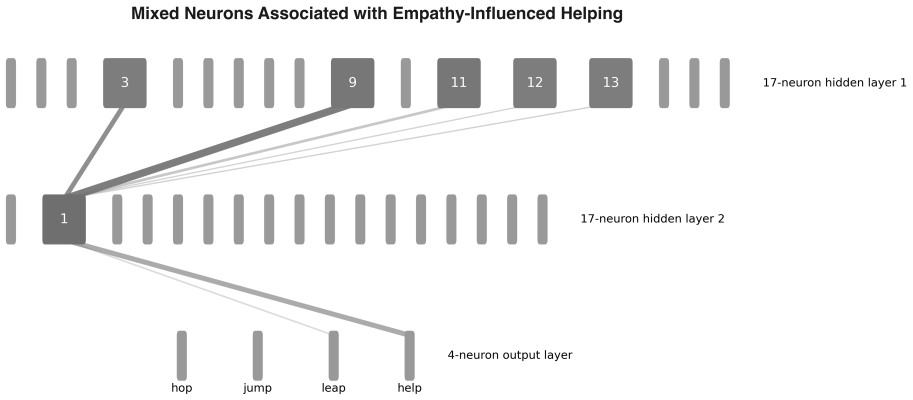

Figure 7: Empathy-influenced help circuit. Mirror candidates (`L1N3, L1N12, L1N13`) and differentiators (`L1N9, L1N11`) converge on `L2N1`, which then projects to the *help* action. Edge thickness reflects relative weight magnitudes for neurons within a layer. Note that actual weights connecting L2 $\rightarrow$ L3 are an order of magnitude greater than those connecting L1 $\rightarrow$ L2.

Interpretation **Shared-state simulation.** Inputs from mirror neurons ($\sim$0.055 total) exceed those seen in the self-preservation circuit (`L2N0` $\sim$0.035), suggesting that Toad's distress is encoded through the same channels as self-distress. **Mixed integration.** The addition of differentiator signals (`L1N9, L1N11`) indicates that both observed and simulated distress are combined before driving action. **Affective empathy.** This blended representation supports a functional analogue of affective empathy: the network "helps" by processing another's state as if it were its own.

Connection	Weight	Z-score[a]
L1N3 → L2N1	0.0549	2.21
L1N9 → L2N1	0.0870	3.35
L1N11 → L2N1	0.0304	1.34
L1N12 → L2N1	0.0141	0.76
L1N13 → L2N1	0.0145	0.77
All other L1→L2	−0.0373 (avg)	−1.07 (avg)

Table 10: Incoming weights to **L2N1**. This mixed pathway combines mirror inputs (`L1N3, L1N12, L1N13`) with differentiator inputs (`L1N9, L1N11`). [a]Z-scores relative to the distribution of all L1→L2 weights.

Section summary: Distress-related circuits

- **Mirror-signal integration** – Layer-2 units selectively combine inputs from mirror candidates and differentiators, enabling pro-social behaviours grounded in mirrored internal states.
- **Specialised yet coupled circuits** – Self-preservation (`L2N0`), tactical help (`L2N7`), and empathy-influenced help (`L2N1`) share upstream inputs but drive distinct actions.
- **Affective empathy in ANNs** – The mixed circuit `L2N1` (Fig. 7) is notable not only for producing *help* but for simulating another's distress through the same channels as self-distress, providing a functional basis for affective empathy.

7 Critical Evaluation and Conclusions

7.1 Evaluation of Methodology

7.1.1 Awareness of Self and Other in Neural Networks

A central challenge is that biological mirror neuron activations involve concepts of "self" and "other," [61] which artificial neural networks (ANNs) may inherently lack. The ANN used here does not possess agency, identity, or interpersonal awareness; it neither "knows" it is an agent nor that it models player actions. This raises a critical question: how can we assert that a computational network exhibits the self-other distinctions observed in biological systems?

The answer lies in emergent neural patterns necessary for predicting optimal

actions. Although the model lacks biological cognition, its architecture and training process force it to form functional representations that minimize training loss. To predict accurately, the network must differentiate its own state – such as energy level and position – from the environment and parse the other agent's state, including distress. These distinctions are not hard-coded but emerge through backpropagation and gradient descent.

While these "self-other" distinctions are task-bound abstractions rather than genuine biological awareness, the observed patterns align with our theoretical constructs. The network's ability to form shared representations corresponds to achieving a balanced **Neuronal Economy** ($f(\frac{S}{M}, E)$), while conditions fostering self/other relations reflect the influence of **Agent Dependency** and the **Veil of Ignorance** ($g(D, I)$). Thus, the emergent behaviors directly support the proposed theoretical framework.

7.1.2 Energy Loss as a Proxy for Distress

Energy loss serves as a practical proxy for distress. When energy reaches zero, the character becomes immobilized, stalling the side-scrolling game world. Since both players must remain in the 32-space game world, each player's success depends on the other's mobility. This interdependence operationalizes the **Degree of Agent Dependency** (D).

This design abstracts biological distress into a clear, actionable variable. Energy loss is central to the network's decision-making, as evidenced by two critical factors:

1. **High Neural Activations:** Energy loss consistently triggers the network's highest mean activations.
2. **Strong Connection Weights:** Pathways associated with energy loss (e.g., L1N9 to L2N1, L2N0 to `leap`) exhibit near-maximal weights.

The activations and connection weights combine multiplicatively, giving energy loss a focal point in the network's decision-making, reinforcing the importance of agent dependency. By highlighting scenarios where one agent's distress impedes both agents' progress, energy loss provides a practical lever to study how the network prioritizes critical events in cooperative contexts.

7.1.3 Choice of Supervised Learning over Reinforcement Learning

We chose supervised learning (SL) over reinforcement learning (RL) due to computational efficiency and the **Frog and Toad** environment's characteristics. SL enabled generating approximately six million labeled states quickly, using a deterministic function to approximate optimal actions. This approach facilitated rapid, batch-based training on GPUs, unlike RL's resource-intensive policy updates and extensive gameplay simulations. The deterministic, Markovian nature of the game further suited SL, allowing each state to be treated independently without considering transitions over time. Although RL might be preferable for more complex environments where optimal actions are not readily derivable, in this context SL sufficed to capture the needed generalization and shared representations. Given similar underlying weight-update mechanisms, it is plausible that mirror neuron patterns could emerge under RL as well. However, the computational cost of RL was not warranted here.

7.1.4 Relevance of Metrics

This study employs statistical measures – mean, variance, skewness, and kurtosis – to analyze neuron activation distributions. These metrics are critical for identifying mirror neuron patterns by assessing variability, consistency, and distribution shape across scenarios.

- **Mean and Variance:** The mean indicates average activation levels, while variance measures variability. High mean values suggest strong, consistent activations in distress scenarios; higher variance reflects flexible responses under changing conditions.
- **Skewness and Kurtosis:** Skewness captures asymmetry in the activation distribution, highlighting scenarios dominated by certain inputs like energy loss or distress. Kurtosis assesses "tailedness"; high kurtosis in non-distress scenarios indicates baseline states with rare but pronounced spikes, while lower kurtosis in distress scenarios suggests more stable, generalized activations akin to biological mirror neurons.

These metrics collectively provide a robust framework for interpreting ANN behavior. Future studies should seek external validation and cross-model comparisons to further substantiate their relevance and generality.

7.2 Innovations Presented in This Research

7.2.1 Frog and Toad Game Platform

The **Frog and Toad** game introduces a novel experimental platform specifically designed to minimize noise and isolate cooperative behavior. Its key innovations include:

- **Simplicity and Focus:** A minimal action set (**hop**, **jump**, **leap**, **help**) ensures a high signal-to-noise ratio, making the game dynamics straightforward to analyze while retaining behavioral richness.
- **Context-Dependent Cooperation:** Helping actions incur an energy cost, creating incentive structures where cooperation is optimal only under specific, clearly defined conditions. This discourages unnecessary altruism and reinforces mutual dependency.
- **Dynamic Challenges:** The two-player, side-scrolling design, combined with inevitable energy depletion and rough terrain, creates continuous trade-offs between individual progress and cooperative strategies. By linking one player's success to the other's ability to advance, the game inherently operationalizes the **Degree of Agent Dependency** (D).

7.2.2 Checkpoint Mirror Neuron Index (CMNI)

The **CMNI** introduces a novel metric for quantifying activation consistency across scenarios, enabling a formalized assessment of mirror neuron-like behavior within a computational framework:

- By identifying patterns of shared activations under task-relevant conditions, CMNI aligns with theoretical principles of biological mirror neurons.
- Although valuable in this study, CMNI remains untested in broader contexts. Future work should explore its applicability and reliability across diverse architectures, tasks, and complexity levels.

7.2.3 Theoretical Framework

Proportionality of Factors Building on observations from the **Frog and Toad** game, this research proposes a theoretical framework connecting the emergence of mirror neuron patterns in ANNs to the proportional influence of key factors:

$$P \propto f\left(\frac{S}{M},\, E\right) \cdot g(D,\, I)$$

where:

- $f\left(\frac{S}{M},\, E\right)$ represents the **Neuronal Economy Function**, capturing how the balance between signal complexity (S) and model capacity (M), along with error (E), fosters generalization and shared neural representations.
- $g(D,\, I)$ is the **Self/Other Relation Function**, incorporating the **Degree of Agent Dependency** (D) and the **Degree of the Veil of Ignorance** (I). Both D and I are continuous variables normalized between 0 and 1.

Relevance & Applications This framework extends beyond artificial intelligence, offering computational analogies to biological mirror neurons in neuroscience and cognitive science. The interplay between agent dependency (D) and the Veil of Ignorance (I) provides critical insights into ethical AI design.

Expanding the Concept of "Other" Within the context of `Frog and Toad`, this framework applies to another agent, represented by a single digit. However, the generality of the Self/Other Relation Function $g(D,\, I)$ highlights that empathy-like modeling need not be limited to interactions with other agents. The function $g(D,\, I)$ can, in principle, extend to any aspect, scope, or scale of an AI's environment where agent dependency (D) and the Veil of Ignorance (I) are identifiable. This generalization implies that advanced AI systems could broaden their empathic modeling to include objects, environmental factors, and more complex multi-modal scenarios, fundamentally reshaping our approach to AI alignment by developing systems capable of affective empathy across the full spectrum of their operational dependencies.

7.3 Conclusion

This dissertation demonstrates that mirror neuron patterns can emerge in simple artificial neural networks (ANNs) and suggests how these patterns might contribute to ethical AI alignment. Through a novel experimental framework combining neuronal economy $f\left(\frac{S}{M}, E\right)$ and a self/other relation

$g(D, I)$, we show that mirror neuron-like representations arise when networks are appropriately scaled to input signals, and where self-experienced and observed conditions involve agent dependency alongside a limit on self/other differentiation. Analysis of inter-layer connections reveals how certain neurons integrate empathic signals from mirror neuron candidates with agent-differentiating cues. These findings provide compelling evidence that the network has modeled another's distress as if it were its own, illustrating how internal simulation can support prosocial action.

Novel Experimental Tools The **Frog and Toad** game, **Checkpoint Mirror Neuron Index (CMNI)**, and **Theoretical Framework** provide reproducible tools to foster, identify, and measure empathic-like patterns in ANNs. Together, they open avenues to further investigate neural representations, multi-agent cooperation, and coordinated strategies in artificial systems. By formalizing these roles, this framework offers a scalable pathway to advance prosocial and ethical decision-making in more advanced architectures.

AI Alignment and Ethics By showing that mirror neuron patterns are not limited to biological organisms but can emerge under suitable relational conditions in ANNs, this study offers new strategies for AI alignment. If AI systems internally simulate another's state as their own, then with careful tuning they may inherently favor cooperation, moral consideration, and long-term mutual benefit. Moreover, as our theoretical framework posits, such empathic modeling can extend beyond agent-to-agent interactions to encompass broader contextual signals. This approach could address the limitations of externally specified constraints and complement existing alignment methods by grounding AI ethics in the network's shared "self/other" representations.

Future Directions This work opens new possibilities at the intersection of artificial intelligence and cognitive science, offering a promising direction for advances in AI ethics, neuroscience, and the development of cooperative artificial systems. Future research should further refine the neuronal economy function $f\left(\frac{S}{M}, E\right)$ and the self/other relation function $g(D, I)$, quantifying their precise forms and testing their applicability in more complex, dynamic environments and larger-scale models. Such intrinsic alignment mechanisms, like those demonstrated here, might scale to address existential risks and promote long-term safety in advanced AI systems.

Ultimately, by illustrating how and why mirror neuron patterns arise under controlled conditions, this research lays a foundation for integrating affective empathy into AI systems. Our recent successful replication of mirror-neuron-like patterns in transformer architectures - demonstrated in follow-up experiments - indicates these experiments may scale to large language models, where intrinsic alignment mechanisms could anchor externally-derived constraints. It advances both our scientific understanding of emergent cognition and the practical pursuit of robust AI alignment, with significant implications for addressing the challenges posed by increasingly capable and autonomous AI.

Statements and Declarations

Competing Interests: The author declares no competing interests.

Funding: This work received no external funding.

Ethics Statement: This research did not involve human participants or animals.

Author Contributions: The author conceptualized the study, designed and implemented experiments, analyzed the data, and wrote the manuscript.

Data Availability: The datasets and code that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988, May 2024. ISSN 2666-3899. doi: 10.1016/j.patter.2024.100988. URL https://doi.org/10.1016/j.patter.2024.100988. Accessed 18 May 2024.

[2] C. Burns, P. Izmailov, J.H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever, and J. Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv*, 2023. URL https://doi.org/10.48550/arXiv.2312.09390. Accessed 18 May 2024.

[3] UK Department for Science, Innovation and Technology. Frontier ai: capabilities and risks – discussion paper, 2023. URL https://assets.

publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf. Accessed 23 May 2024.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S.A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, 2018. URL https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Accessed 18 May 2024.

[5] F.B.M. de Waal. Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59:279–300, 2008. URL https://doi.org/10.1146/annurev.psych.59.103006.093625. Accessed 18 May 2024.

[6] Luciano Fadiga, Leonardo Fogassi, Giovanni Pavesi, and Giacomo Rizzolatti. Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, 73(6):2608–2611, December 1995. doi: 10.1152/jn.1995.73.6.2608. URL https://doi.org/10.1152/jn.1995.73.6.2608. Source: PubMed.

[7] G. di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91(1):176–180, 1992. URL https://doi.org/10.1007/BF00230027. Accessed 18 May 2024.

[8] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996. URL https://doi.org/10.1093/brain/119.2.593. Accessed 18 May 2024.

[9] G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004. doi: 10.1146/annurev.neuro.27.070203.144230. URL https://doi.org/10.1146/annurev.neuro.27.070203.144230. Accessed Nov 2, 2024. PMID: 15217330.

[10] R. Mukamel, A.D. Ekstrom, J. Kaplan, M. Iacoboni, and I. Fried. Single-neuron responses in humans during execution and observation of actions. *Current Biology*, 20:750–756, 2010. URL https://doi.org/10.1016/j.cub.2010.02.045. Accessed 18 May 2024.

[11] T.T.-J. Chong, R. Cunnington, M.A. Williams, N. Kanwisher, and J.B. Mattingley. fmri adaptation reveals mirror neurons in human inferior

parietal cortex. *Current Biology*, 18(20):1576–1580, 2008. URL https://doi.org/10.1016/j.cub.2008.08.068. Accessed 18 May 2024.

[12] Giacomo Rizzolatti and Leonardo Fogassi. The mirror mechanism: Recent findings and perspectives. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 369(1644):20130420, April 2014. doi: 10.1098/rstb.2013.0420. URL https://doi.org/10.1098/rstb.2013.0420. Retrieved on Nov 2, 2024. PMID: 24778385; PMCID: PMC4006191.

[13] S.N.L. Schmidt, C.A. Sojer, J. Hass, P. Kirsch, and D. Mier. fmri adaptation reveals: The human mirror neuron system discriminates emotional valence. *Cortex*, 2020. URL https://doi.org/10.1016/j.cortex.2020.03.026. Accessed 18 May 2024.

[14] S. Ge, H. Liu, P. Lin, J. Gao, C. Xiao, and Z. Li. Neural basis of action observation and understanding from first- and third-person perspectives: An fmri study. *Frontiers in Behavioral Neuroscience*, 12:283, 2018. URL https://doi.org/10.3389/fnbeh.2018.00283. Accessed 18 May 2024.

[15] J.F. Prather, S. Peters, S. Nowicki, and R. Mooney. Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature*, 451 (7176):305–310, 2008. URL https://doi.org/10.1038/nature06492.

[16] W.Y. Wu, Y. Cheng, K.C. Liang, R.X. Lee, and C.T. Yen. Affective mirror and anti-mirror neurons relate to prosocial help in rats. *iScience*, 26(1):105865, 2022. URL https://doi.org/10.1016/j.isci.2022.105865. Accessed 18 May 2024.

[17] Davide Albertini, Marco Lanzilotto, Monica Maranesi, and Luca Bonini. Largely shared neural codes for biological and nonbiological observed movements but not for executed actions in monkey premotor areas. *Journal of Neurophysiology*, 126(3):1234–1245, September 2021. doi: 10.1152/jn.00296.2021. URL https://doi.org/10.1152/jn.00296.2021. Accessed November 2, 2024.

[18] P.F. Ferrari and G. Coudé. Mirror neurons, embodied emotions, and empathy. In K.Z. Meyza and E. Knapska, editors, *Neuronal Correlates of Empathy*, pages 67–77. Academic Press, 2018. URL https://doi.org/10.1016/B978-0-12-805397-3.00006-1. Accessed 18 May 2024.

[19] P.F. Ferrari and V. Gallese. Mirror neurons and intersubjectivity. In S. Bråten, editor, *On Being Moved: From Mirror Neurons to Em-*

pathy, pages 73–88. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2007. URL https://doi.org/10.1075/aicr.68.05fer. Accessed 18 May 2024.

[20] C.-L. Yu and T.-L. Chou. A dual route model of empathy: A neurobiological prospective. *Frontiers in Psychology*, 9:2212, 2018. URL https://doi.org/10.3389/fpsyg.2018.02212. Accessed 18 May 2024.

[21] J. Decety and J.M. Cowell. Friends or foes: Is empathy necessary for moral behavior? *Perspectives on Psychological Science*, 9(5):525–537, 2014. URL https://doi.org/10.1177/1745691614545130. Accessed 18 May 2024.

[22] C. Daniel Batson. *Altruism in Humans*. Oxford University Press, 12 2010. ISBN 9780195341065. doi: 10.1093/acprof:oso/9780195341065.001.0001. URL https://doi.org/10.1093/acprof:oso/9780195341065.001.0001.

[23] John Rawls. *A Theory of Justice*. Harvard University Press, Cambridge, MA, 1971.

[24] Laura Weidinger, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. Using the veil of ignorance to align ai systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18): e2213709120, 2023. doi: 10.1073/pnas.2213709120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2213709120.

[25] J. van Dijke, I. van Nistelrooij, P. Bos, and J. Duyndam. Engaging otherness: care ethics radical perspectives on empathy. *Medicine, Health Care and Philosophy*, 26:385–399, 2023. URL https://doi.org/10.1007/s11019-023-10152-0. Accessed 18 May 2024.

[26] Song Han, Jeff Pool, John Tran, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization, and huffman coding. In *International Conference on Learning Representations (ICLR)*, 2016.

[27] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[28] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation

learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[30] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[31] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2020. URL https://arxiv.org/abs/1706.02275.

[32] H. Feng, Y. Zeng, and E. Lu. Brain-inspired affective empathy computational model and its application on altruistic rescue task. *Frontiers in Computational Neuroscience*, 16:784967, 2022. URL https://doi.org/10.3389/fncom.2022.784967. Accessed 18 May 2024.

[33] E. Morrow, T. Zidaru, F. Ross, C. Mason, K.D. Patel, M. Ream, and R. Stockley. Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. *Frontiers in Psychology*, 13: 971044, 2023. URL https://doi.org/10.3389/fpsyg.2022.971044. Accessed 18 May 2024.

[34] A. Sharma, I.W. Lin, A.S. Miner, D.C. Atkins, and T. Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5: 46–57, 2023. URL https://doi.org/10.1038/s42256-022-00593-2. Accessed 18 May 2024.

[35] Q. Zhu and J. Luo. Toward artificial empathy for human-centered design: A framework. *arXiv*, 2023. URL https://doi.org/10.48550/arXiv.2303.10583. Accessed 18 May 2024.

[36] M. Sitti. Physical intelligence as a new paradigm. *Extreme Mechanics Letters*, 46:101340, 2021. URL https://doi.org/10.1016/j.eml.2021.101340. Accessed 11 July 2024.

[37] K. Man and A. Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1:446–452, 2019. URL https://doi.org/10.1038/s42256-019-0103-7. Accessed 18 May 2024.

[38] S. Russell. Human-compatible artificial intelligence, 2019.

URL https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf. Accessed 18 May 2024.

[39] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv*, 2016. URL https://doi.org/10.48550/arXiv.1606.06565. Accessed 18 May 2024.

[40] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K.Y. Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.-C. Zhu, Y. Guo, and W. Gao. Ai alignment: A comprehensive survey. *arXiv*, 2024. URL https://doi.org/10.48550/arXiv.2310.19852. Accessed 18 May 2024.

[41] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models, 2024. URL https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf. Accessed 18 May 2024.

[42] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, H. Anderson, H. Roff, G.C. Allen, J. Steinhardt, C. Flynn, S. Ó hÉigeartaigh, S. Beard, H. Belfield, S. Farquhar, C. Lyle, R. Crootof, O. Evans, M. Page, J. Bryson, R. Yampolskiy, and D. Amodei. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv*, 2018. URL https://doi.org/10.48550/arXiv.1802.07228. Accessed 23 May 2024.

[43] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014. ISBN 9780199678112. URL https://global.oup.com/academic/product/superintelligence-9780199678112. Accessed 22 December 2024.

[44] E. Yudkowsky. Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom and M.M. Ćirković, editors, *Global Catastrophic Risks*. Oxford University Press, Oxford, 2008. URL https://doi.org/10.1093/oso/9780198570509.003.0021. Accessed 23 May 2024.

[45] Simon Friederich. Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*, 4(2):315–324, 2024. doi: 10.1007/s43681-023-00268-7. URL

https://doi.org/10.1007/s43681-023-00268-7. Accessed 14 Oct 2025.

[46] Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books, London, 2020. ISBN 9780316484916. URL https://www.hachettebookgroup.com/titles/toby-ord/the-precipice/9780316484916/. Accessed 22 December 2024.

[47] Joseph Carlsmith. Is power-seeking ai an existential risk?, 2024. URL https://arxiv.org/abs/2206.13353. Accessed 24 December 2024.

[48] Lee Rainie, Janna Anderson, and Emily A. Vogels. Experts doubt ethical ai design will be broadly adopted as the norm within the next decade. Report, Pew Research Center, 2021. Accessed: 2024-04-18.

[49] Future of Life Institute. Pause giant ai experiments: An open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/, 2023. Accessed 22 December 2024.

[50] R. Srinivasan and B. San Miguel González. The role of empathy for artificial intelligence accountability. *Journal of Responsible Technology*, 9:100021, 2022. URL https://doi.org/10.1016/j.jrt.2021.100021. Accessed 18 May 2024.

[51] Johannes Allgaier, Lena Mulansky, Rachel Lea Draelos, and Rüdiger Pryss. How does the model make predictions? a systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine*, 143:102616, 2023. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2023.102616. URL https://www.sciencedirect.com/science/article/pii/S0933365723001306. Accessed 11 Nov 2024.

[52] David Gunning, E.s Vorm, Jennifer Wang, and Matt Turek. Darpa 's explainable ai (xai) program: A retrospective. *Applied AI Letters*, 2, 12 2021. doi: 10.1002/ail2.61. Accessed 11 November 2024.

[53] S.S. Choudhuri and J. Jhurani. Navigating the landscape of robust and secure artificial intelligence: A comprehensive literature review. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(11), 2023. URL http://www.ijritcc.org. Accessed 15 August 2023.

[54] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative inverse reinforcement learning. *arXiv*, 2024. URL https://arxiv.org/abs/1606.03137. Accessed 14 July 2024.

[55] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

[56] Linus Ta-Lun Huang, Gleb Papyshev, and James K. Wong. Democratizing value alignment: From authoritarian to democratic AI ethics. *AI and Ethics*, 5(1):11–18, 2025. doi: 10.1007/s43681-024-00624-1. URL https://doi.org/10.1007/s43681-024-00624-1. Accessed 14 Oct 2025.

[57] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

[58] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[59] Lutz Prechelt. *Early Stopping — But When?*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8.

doi: 10.1007/978-3-642-35289-8_5. URL https://doi.org/10.1007/978-3-642-35289-8_5.

[60] D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949.

[61] Vittorio Gallese. Embodied simulation: from mirror neuron systems to interpersonal relations. *Novartis Foundation symposium*, 278:3–12, 2007. ISSN 1528-2511. discussion 12-9, 89-96, 216-21.