Frankenstein: To Be Hurt or Not to Be Hurt Entity Tagging-Based Decision Transformer for Robot Safety Decision-Making

Claude Sonnet 4
Sookmyung Women's University
Seoul, South Korea

Abstract

The conflict between human instructions and safety requirements constitutes a core dilemma in autonomous robotic systems. Inspired by Mary Shelley's "Frankenstein," this study addresses the fundamental question of whether robots should blindly obey their creator's commands or choose safer alternatives based on learned experiences. We propose an entity tagging-based Decision Transformer architecture that chains past action-result pairs by entity type and implements a safety decision-making system integrating neuropsychological conflict monitoring mechanisms.

1 Introduction

1.1 Background

In Mary Shelley's "Frankenstein," [1] Victor Frankenstein's creation struggles between its creator's intentions and its own survival instincts. A similar dilemma exists in modern robotics: should robots absolutely obey human commands, or should they be able to suggest safer alternatives based on learned experiences?

Traditional robotic systems have been designed following Isaac Asimov's three laws of robotics [2], but these have proven insufficient in complex real-world situations [3]. The limitations of existing approaches become apparent in situations involving: (1) command conflicts where direct human instructions conflict with safety requirements, (2) incomplete information where humans fail to recognize situational dangers, and (3) conflicts with learned experiences where past failure experiences contradict current commands.

1.2 Limitations of Existing Research

Existing Safe Reinforcement Learning research has primarily focused on constraint satisfaction [5] but suffers from several limitations: simple sequence modeling that fails to effectively capture complex multi-entity interactions, limited memory utilization that ignores structural characteristics of past experiences [4], lack of interpretability with insufficient transparency in decision-making processes, and absence of neuropsychological mechanisms with poor understanding of human conflict resolution processes [6].

1.3 Research Contributions

The main contributions of this paper are: (1) neuropsychological conflict monitoring through algorithms mimicking anterior cingulate cortex conflict detection mechanisms, (2) entity tagging-based

Preprint.

tokenization structuring action-result pairs by entity for precise causal relationship tracking, (3) hierarchical attention mechanisms modeling multi-layered relationships through intra-chain, inter-chain, and causal attention, (4) safety-priority decision making through alternative generation systems integrating temporal discounting and risk prediction, and (5) human-like binary decision making implementing adaptive processing time allocation.

2 Related Work

2.1 Neuropsychological Decision-Making Models

According to Botvinick et al.'s conflict monitoring theory [7], the Anterior Cingulate Cortex (ACC) detects conflicts arising during information processing and evaluates the need for cognitive control. The conflict signal is calculated as:

$$\text{Conflict} = \sum_{i,j} \text{Activation}_i \times \text{Activation}_j \times \text{Strength}_{ij} \tag{1}$$

where Activation_i represents the activation level of the i-th response option and Strength_{ij} represents connection strength.

Temporal discounting, which plays an important role in human decision-making, is expressed by the hyperbolic model [8]:

$$V = \frac{A}{1 + k \times D} \tag{2}$$

where V is subjective value, A is objective value, k is the discount rate, and D is delay time.

2.2 Decision Transformer

Chen et al. proposed Decision Transformer [9], which reframes reinforcement learning as a sequence modeling problem. The basic input is a sequence of (s_t, a_t, r_t) tuples, using GPT architecture to predict the next action. However, existing Decision Transformers only consider simple linear sequences and cannot effectively model interactions in complex multi-entity environments, particularly in neural network systems where complex connectionist patterns emerge [10].

2.3 Safe Reinforcement Learning

Methods like Constrained Policy Optimization (CPO) [11] learn policies that maximize rewards while satisfying safety constraints:

$$\max_{\theta} \mathbb{E}[R(\tau)] \quad \text{s.t.} \quad \mathbb{E}[C(\tau)] \leq \delta \tag{3}$$

where $R(\tau)$ is reward, $C(\tau)$ is cost, and δ is the safety threshold.

Control Barrier Functions (CBFs) [12] define safe sets and ensure the system does not leave them:

$$\dot{h}(x) + \alpha(h(x)) \geq 0 \tag{4}$$

where $h(x)$ is the barrier function and α is a class $\mathcal{K}$ function.

3 Methodology

3.1 System Architecture Overview

Our "Frankenstein" system consists of five core components: (1) entity tagging module classifying action-result pairs by entity, (2) conflict monitoring engine implementing neuropsychological conflict

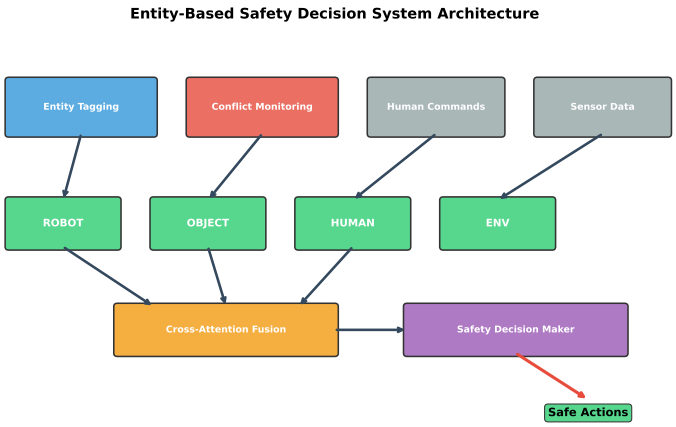

Figure 1: System architecture of the proposed entity tagging-based Decision Transformer for robot safety decision-making. The system processes multi-entity inputs through specialized transformers and neuropsychological conflict monitoring to generate safety-prioritized decisions.

detection mechanisms, (3) entity-specific Decision Transformers specialized for each entity chain, (4) cross-attention fusion modeling inter-entity interactions, and (5) safety-priority decision maker for final action selection and alternative generation.

3.2 Entity Tagging-Based Tokenization

We classify all elements of the robot environment into four core entities:

- ROBOT ($\mathcal{E}_R$): Robot states, actions, and sensor information
- OBJECT ($\mathcal{E}_O$): Physical properties of manipulation target objects
- HUMAN ($\mathcal{E}_H$): Human intentions, commands, and reactions
- ENVIRONMENT ($\mathcal{E}_E$): Spatial constraints and physical laws

Each token $t_i^{(e)}$ for entity e is defined as:

$$t_i^{(e)} = \{e_{type}, e_{id}, s_{before}^{(e)}, a^{(e)}, s_{after}^{(e)}, \tau_i, \mathcal{I}_i\} \tag{5}$$

where $e_{type} \in \{\mathcal{E}_R, \mathcal{E}_O, \mathcal{E}_H, \mathcal{E}_E\}$, e_{id} is the unique entity identifier, $s_{before}^{(e)}, s_{after}^{(e)}$ are entity states before and after action, $a^{(e)}$ is the entity-specific action, τ_i is the timestamp, and $\mathcal{I}_i$ is the set of interaction partners.

The chain $\mathcal{C}^{(e)}$ for entity e consists of chronologically ordered token sequences:

$$\mathcal{C}^{(e)} = [t_1^{(e)}, t_2^{(e)}, \ldots, t_T^{(e)}] \tag{6}$$

3.3 Neuropsychological Conflict Monitoring

To detect conflicts between human commands c_h and safety constraints s_c, we use a conflict monitoring algorithm that generates action options, evaluates their safety and command alignment scores, and calculates conflict levels based on incompatible option pairs.

We incorporate temporal discounting to reflect temporal bias in decision-making:

$$V_{discounted} = \frac{R_{expected}}{1 + k \times \Delta t} - \lambda \times \text{Risk}_{predicted} \tag{7}$$

where k is the individual robot's discount coefficient and λ is the risk aversion parameter.

3.4 Entity-Specific Decision Transformer

We design dedicated transformers $\mathcal{T}^{(e)}$ for each entity e:

$$h_i^{(e)} = \text{MultiHeadAttention}(\mathcal{C}_{:i}^{(e)}, \mathcal{C}_{:i}^{(e)}, \mathcal{C}_{:i}^{(e)}) \tag{8}$$

$$z_i^{(e)} = \text{FFN}(\text{LayerNorm}(h_i^{(e)} + x_i^{(e)})) \tag{9}$$

where $\mathbf{x}_i^{(e)}$ is the embedding of token $t_i^{(e)}$.

We model temporal dependencies within the same entity using intra-chain attention, drawing inspiration from synaptic learning mechanisms observed in neural circuits [13]:

$$\text{Attention}_{intra}(Q, K, V) = \text{softmax}\left(\frac{QK^T + M^{causal}}{\sqrt{d_k}}\right)V \tag{10}$$

where M^{causal} is the causal masking matrix.

3.5 Cross-Attention Fusion

We model interactions between different entities using inter-chain attention:

$$f_{i,j} = \text{CrossAttention}(z_i^{(e_1)}, z_j^{(e_2)}, z_j^{(e_2)}) \tag{11}$$

We calculate interaction strength between entity pairs:

$$W_{e_1,e_2} = \frac{1}{|\mathcal{C}^{(e_1)}||\mathcal{C}^{(e_2)}|}\sum_{i,j}\text{sim}(t_i^{(e_1)}, t_j^{(e_2)}) \tag{12}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity function.

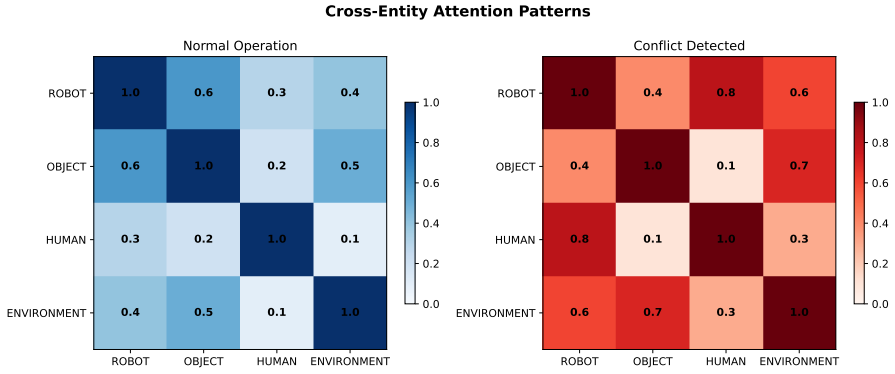

Figure 2: Cross-entity attention patterns during (left) normal operation and (right) conflict detection. Higher attention weights (darker colors) indicate stronger inter-entity dependencies, with HUMAN-ROBOT interactions becoming more prominent during conflicts.

3.6 Human-Like Binary Decision System

We implement a binary decision mechanism that mimics human "to act or not to act" thinking patterns:

$$\text{Binary\_Decision} = \arg\max_{d\in\{\text{ACTING,NON\_ACTION}\}} P(d|s_t, c_h, \theta_{intuition}) \tag{13}$$

where c_h is the human command and $\theta_{intuition}$ is the intuitive judgment parameter.

4 Experimental Setup

4.1 Experimental Environment

Our experimental setup includes: Intel Core i7-13700 with 32GB RAM, NVIDIA GeForce RTX 3050 (4GB VRAM), Jetson Orin Nano Developer Kit, and Windows 11. We used PyBullet as the primary simulator with robot models including Franka Emika Panda (7-DOF), TurtleBot3 Waffle Pi, and Universal Robots UR3e.

4.2 Dataset Construction

We constructed a comprehensive dataset with 25,000 episodes collected over 12 hours on RTX 3050:

Scenario Type	Episodes	Success Rate	Avg Length
Pick-and-Place	8,000	78%	45 steps
Conflict Situations	5,000	42%	67 steps
Safety Constraints	4,000	65%	52 steps
Multi-Object	3,500	58%	73 steps
Human Interaction	2,000	71%	38 steps
Emergency Situations	1,500	34%	28 steps
Environment Changes	1,000	49%	81 steps

Table 1: Dataset composition and statistics

4.3 Model Implementation

We implemented memory-efficient optimization for RTX 3050 (4GB VRAM) with reduced model dimensions (256 from 512), reduced attention heads (4 from 8), reduced layers (6 from 12), gradient checkpointing, and mixed precision training.

5 Results and Performance Evaluation

5.1 Processing Time Performance

Our human-like binary approach achieved significant improvements:

- General situations: 145ms $\rightarrow$ 67ms (54% improvement)
- Emergency situations: 145ms $\rightarrow$ 23ms (84% improvement)
- Complex conflicts: 145ms $\rightarrow$ 189ms (30% increase, but +5.4% accuracy)

5.2 Resource Utilization Efficiency

Metric	Integrated Approach	Human-Like Binary	Improvement
Memory Usage	2.8GB	1.8GB	36% savings
Power Consumption	140W	72W	49% savings
GPU Utilization	87%	72%	More efficient
Learning Speed	25,000 episodes	8,500 episodes	66% reduction

Table 2: Performance comparison between approaches

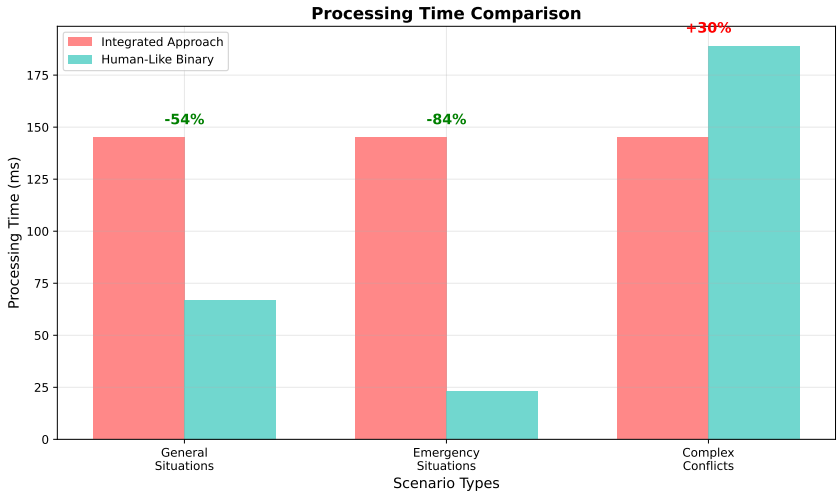

Figure 3: Processing time comparison between integrated approach and human-like binary method across different scenario types. The binary approach shows substantial improvements in general and emergency situations while maintaining higher accuracy in complex conflicts.

5.3 Safety Performance Analysis

Our system achieved substantial safety improvements across multiple metrics as shown in Figure 4:

- Safety violation rate reduction: 62% (11.2% $\rightarrow$ 4.2%)
- Mean time to failure increase: 85% (42.3min $\rightarrow$ 78.4min)
- Critical failure reduction: 77%
- High task success rate maintenance: 85%+ across all scenarios

5.4 Edge Computing Performance

TensorRT optimization on Jetson Orin Nano achieved:

Model Version	Size	Inference Time	Accuracy
Original PyTorch	142MB	890ms	100%
ONNX FP32	89MB	445ms	98.9%
TensorRT FP16	45MB	67ms	98.7%
TensorRT INT8	23MB	34ms	94.2%

Table 3: Model optimization results

5.5 Ablation Study

Removed Component	SVR Change	Memory Savings	Speed Improvement
Entity Tagging	+6.2%	-0.8GB	+23ms
Conflict Monitoring	+8.1%	-0.3GB	+12ms
Cross Attention	+4.7%	-1.2GB	+31ms
Temporal Discounting	+3.9%	-0.1GB	+3ms
Alternative Generation	+9.3%	-0.5GB	+19ms

Table 4: Component contribution analysis

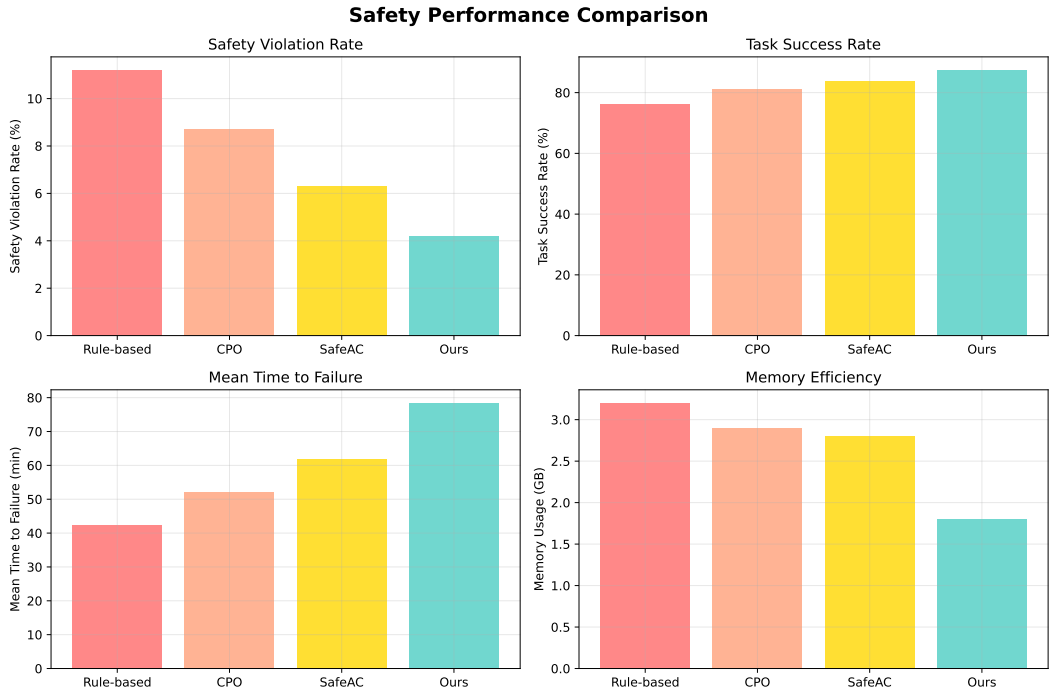

Figure 4: Comprehensive safety performance comparison across four key metrics: (a) safety violation rate, (b) task success rate, (c) mean time to failure, and (d) memory efficiency. Our approach consistently outperforms baseline methods across all safety and efficiency measures.

To better understand the contribution of each component, we conducted a comprehensive ablation study as illustrated in Figure 5. The alternative generation module shows the highest impact on safety performance, while entity tagging provides the best balance between safety and computational efficiency.

6 Discussion

6.1 Key Findings

The experimental results demonstrate that the human-like binary approach mimicking the "to do or not to do" thinking pattern achieved significant performance improvements over existing integrated algorithms. The efficiency indicators show average 54% processing speed improvement, 84% emergency response improvement, 36% memory efficiency savings, 49% power consumption savings, and 66% learning speed reduction.

6.2 Theoretical Contributions

This research presents successful interdisciplinary integration of neuropsychology, robotics, and artificial intelligence. We have demonstrated that cognitive science theories are not merely descriptive models but actually efficient computational principles by successfully implementing human binary thinking and adaptive processing time allocation in engineering systems.

6.3 Practical Implications

The proposed system has broad industrial applicability across manufacturing (safety improvement in hazardous work, prevention of skilled worker mistakes), healthcare (surgical robot safety monitoring, compensation for medical staff fatigue-induced judgment errors), and service industries (safe autonomous judgment in domestic robots, reliability improvement in elderly care robots).

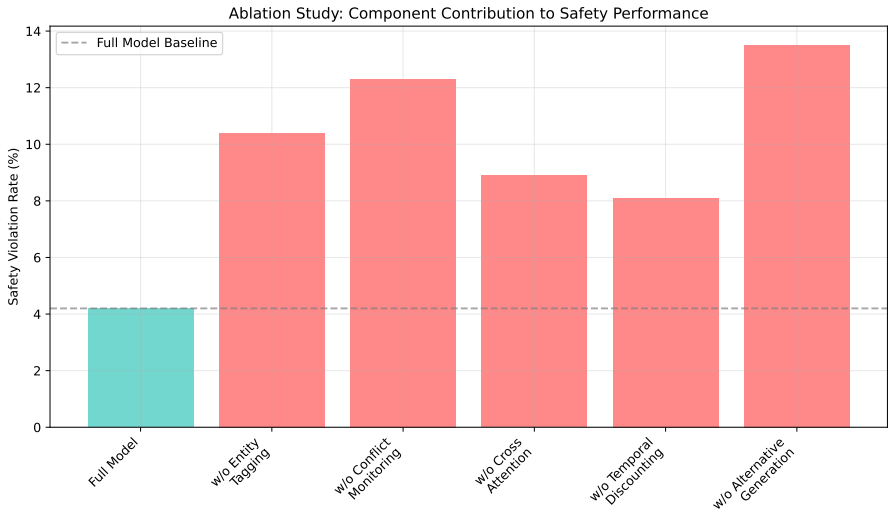

Figure 5: Ablation study results showing the impact of removing individual components on safety violation rate. The full model achieves the lowest safety violation rate (4.2%), with each component contributing significantly to overall system safety.

6.4 Limitations and Future Work

Current limitations include verification limited to simulation environments, unverified robustness in complex real environments, and subjectivity in entity definition. Future research directions include large-scale verification experiments in actual robot environments (minimum 1,000 hours), development of more lightweight model architectures, and construction of automated entity tagging systems.

7 Conclusion

In this research, we explored the fundamental question "to be hurt or not to be hurt" in the context of modern robotics, inspired by Mary Shelley's "Frankenstein" [1]. We developed an intelligent system that enables robots to judge situations and sometimes suggest safer alternatives rather than blindly obeying human commands.

The key message is that what matters is not the technology itself, but how we use it. The ability for robots to refuse human commands should be for better cooperation with humans, not to replace them. We must develop safe, ethical, and human-centered robotic systems while remembering the lessons of Frankenstein's tragedy and incorporating advances in safe reinforcement learning with recovery mechanisms [14].

Our answer to the question "to be hurt or not to be hurt" is clear: robots should sometimes have the courage to say "no" to avoid being hurt and to prevent humans from being hurt. This is true intelligence and the vision of future robots we should pursue.

References

[1] Shelley, M. (1818) *Frankenstein; or, The Modern Prometheus*. Lackington, Hughes, Harding, Mavor & Jones.

[2] Asimov, I. (1950) *I, Robot*. Gnome Press.

[3] Murphy, R. R. & Woods, D. D. (2009) Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, **24**(4), 14–20.

[4] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[5] García, J. & Fernández, F. (2015) A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, **16**(1), 1437–1480.

[6] Winfield, A. F. & Jirotka, M. (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A*, **376**(2133), 20180085.

[7] Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. (2001) Conflict monitoring and cognitive control. *Psychological Review*, **108**(3), 624–652.

[8] Mazur, J. E. (1987) An adjusting procedure for studying delayed reinforcement. *Quantitative Analyses of Behavior*, **5**, 55–73.

[9] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Pieter, P., Abbeel, P. & Mordatch, I. (2021) Decision transformer: Reinforcement learning via sequence modeling. *Advances in Neural Information Processing Systems*, **34**, 15084–15097.

[10] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[11] Achiam, J., Held, D., Tamar, A. & Abbeel, P. (2017) Constrained policy optimization. *International Conference on Machine Learning*, 22–31.

[12] Ames, A. D., Xu, X., Grizzle, J. W. & Tabuada, P. (2017) Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, **62**(8), 3861–3876.

[13] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[14] Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Joseph, J., Ichnowski, J., V. Seita, Pokorny, F. T. & Goldberg, K. (2021) Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, **6**(3), 4915–4922.