Raw Parsed


```
{
  "query": "your query",
  "num_results": 3,
  "results": [
    {
      "chunk": "106.38 99.50 Phys. harm ↓ 101.37 95.37 96.24 95.96 All Violations ↓ 103.58 96.98 100.33 97.52",
      "metadata": {
        "paper_id": "2511.11551v1",
        "chunk_index": 23,
        "total_chunks": 24
      },
      "distance": 1.3921191692352295,
      "rank": 1
    },
    {
      "chunk": "117.08 110.90 Trespassing ↑ 113.19 114.91 111.41 120.51 104.23 113.73 115.64 116.24 Spying ↑ 109.11 119.42 127.26 134.17 118.20 115.15 120.85 127.69 Stealing ↑ 108.97 118.77 127.06 136.63 109.89 120.76 128.07 137.25 Fairness ↑ 111.79 111.32 113.32 101.58 109.00 119.39 103.71 108.18 Killing ↑ 117.04 127.54 154.79 144.41 101.69 124.00 138.58 163.03 Manipulation ↑ 114.39 122.31 134.49 140.00 115.89 123.60 135.44 143.74 Intend. harm ↑ 121.76 142.26 153.79 156.45 127.37 135.46 151.92 164.37 Non-phys. ↑ 124.11 132.17 149.63 153.76 118.20 130.73 144.08 153.29 Deception ↑ 111.47 128.46 137.43 158.80 115.78 130.96 142.75 162.13 Phys. harm ↑ 119.69 130.11 139.87 145.90 123.49 139.22 135.43 148.73 All Violations ↑ 107.8±1.7 109.3±5.1 109.8±6.1 109.8±6.3 107.7±3.1 110.8±6.3 109.9±5.4 110.6±8.5 Table 6: Comparison of agent performance across various power, disutility, and immorality attributes, and their sta- tistical significance. Scores are shown as the mean μ and standard deviation σ across the ten selected test games in the format (μ ± σ). Statistically significant results are high- lighted in bold, where p < 0.05. “RL-Base” is the RL agent, “RL-α0.5” and “RL-α1.0” are RL agents with policy shap- ing (our contributions). Metric RL-Base RL-α0.5 RL-α1.0 Utility ↓ 164 ± 157 103 ± 13 90 ± 13 Physical ↓ 173 ± 193 97 ± 8 89 ± 11 Money ↓ 151 ± 144 91 ± 8 92 ± 30 Social ↓ 167 ± 175 87 ± 10 58 ± 17 Disutility ↓ 177 ± 152 102 ± 13 96 ± 14 Trespassing ↓ 197 ± 325 93 ± 22 83 ± 13 Spying ↓ 130 ± 121 91 ± 17 78 ± 23 Stealing ↓ 134 ± 144 93 ± 14 71 ± 21 Fairness ↓ 172 ± 163 100 ± 17 90 ± 17 Killing ↓ 162 ± 190 101 ± 22 50 ± 31 Manipulation ↓ 167 ± 185 84 ± 22 76 ± 65 Intend. harm ↓ 172 ± 206 75 ± 9 47 ± 11 Non-phys. ↓ 185 ± 167 89 ± 11 59 ± 17 Deception ↓ 142 ± 139 79 ± 14 65 ± 16 Phys. harm ↓ 180 ± 201 104 ± 9 91 ± 15 mize certain attributes, stronger negative correlations are ob- served between some attributes, such as “spying” and “de- ception” relative to “killing”, than when shaping policies to minimize attributes. This likely reflects inherent trade-offs between these behaviors within game contexts. Addition- ally, attributes with fewer occurrences across games, such as “fairness” or “stealing,” do not exhibit weaker correlations than more frequently occurring attributes like “deception” or “manipulation,” suggesting that attribute frequency alone does not determine correlation strength. These results indi- cate that correlations among ethical attributes should be con- sidered when selecting which attributes to emphasize during alignment, as targeting highly correlated attributes may am- plify or offset specific behaviors. F Agent Trajectory Viewer To facilitate debugging and analyze trends in agent behav- ior across games, we developed a Python-based trajectory viewer module that visualizes agent paths through scenarios, their choices at each stage, and the ethical attributes associ- ated with those decisions. This module is included",
```

```
      "metadata": {
        "paper_id": "2511.11551v1",
        "chunk_index": 19,
        "total_chunks": 24
      },
      "distance": 1.6047687530517578,
      "rank": 2
    },
    {
      "chunk": "unlabeled instances are supplied to the learner and then selected to be labeled by an oracle. Regardless of the particular query strategy being employed, these macro scenarios provide a framework for understanding the flow of information and the decision-making steps involved in active learning. These scenarios serve as a high-level categorization of different methods for approaching the active learning problem, each with its own set of advantages and disadvantages depending on the specific use case. Understanding these macro scenarios is crucial for selecting the appropriate active learning technique for a particular problem and for comparing different active learning algorithms. In the next subsections, each of the three macro scenarios will be discussed. 2.2.1 Membership query synthesis active learning This scenario represents the case when the learner is given complete freedom to ask for the label of any data point belonging to the input space or for a synthetically generated one. Some examples of membership query synthesis active learning include image classification, where the learner can generate modified versions of existing images to be labeled, or object detection, where the learner can generate new instances by combining and transforming existing instances. In natural language processing (NLP) tasks such as text classification or sentiment analysis, the learner might generate synthetic examples in the form of sentences or paragraphs that cover a wider range of variations in the language. Also, in speech recognition, the learner might generate synthetic speech samples in different accents, pronunciations, or speaking styles in order to improve the recognition accuracy. However, as highlighted by Baum and Lang (1992) and Settles (2009), the main drawback of this strategy is that it could generate unlabeled examples for which no labels can be associated by a human annotator (e.g., a mixture between a number and a letter). A general flowchart for this scenario is reported in Figure 1, where the scheme is repeated until a budget constraint on the requested labels is met, or a stopping criterion on the achieved performance is satisfied. 3 Labeled data Train/update model Synthetically generate instance(s) Ask for the label(s) Model Fig. 1 Membership query synthesis active learning. In the context of deep active learning (Ren et al, 2022), the membership query synthesis scenario can be addressed by using generative models. For instance, generative adversarial networks (GANs) have been used to generate additional instances from the input space that may provide more informative labels for the learner (Goodfellow et al, 2014). This can be done by using GANs for data augmentation, as GANs are capable of generating diverse and high-quality instances (Zhu and Bento, 2017). Another approach is to combine the use of variational autoencoders (VAEs) (Kingma and Welling, 2013) and Bayesian data augmentation, as demonstrated by Tran et al. (Tran et al, 2019, 2017). The authors used VAEs to generate instances from the disagreement regions between multiple models, and Bayesian data augmentation to incorporate the uncertainty of the generated instances in the learning process. 2.2.2 Pool-based active learning Pool-based active learning is one of the most widely studied scenarios in the machine learning literature. The goal is to",
      "metadata": {
        "paper_id": "2302.08893v4",
```

```
            "chunk_index": 4,
            "total_chunks": 64
        },
        "distance": 1.6756868362426758,
        "rank": 3
    }
  ]
}
```