Principles2Plan: LLM-Guided System for Operationalising Ethical Principles into Plans

Tammy Zhong, Yang Song, Maurice Pagnucco

School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
{tammy.zhong,yang.song1,morri}@unsw.edu.au

Abstract

Ethical awareness is critical for robots operating in human environments, yet existing automated planning tools provide little support. Manually specifying ethical rules is labour-intensive and highly context-specific. We present *Principles2Plan*, an interactive research prototype demonstrating how a human and a Large Language Model (LLM) can collaborate to produce context-sensitive ethical rules and guide automated planning. A domain expert provides the planning domain, problem details, and relevant high-level principles such as beneficence and privacy. The system generates operationalisable ethical rules consistent with these principles, which the user can review, prioritise, and supply to a planner to produce ethically-informed plans. To our knowledge, no prior system supports users in generating principle-grounded rules for classical planning contexts. Principles2Plan showcases the potential of human-LLM collaboration for making ethical automated planning more practical and feasible.

Introduction

The deployment of robots around people raises the challenge of ensuring that their actions achieve goals while respecting ethical principles. High-level ethical principles, such as beneficence, depend heavily on context. For example, in an autonomous vehicle scenario, a passenger needing urgent medical attention may justify taking an unauthorised shortcut to reach the hospital quickly, whereas for a leisure trip, following standard traffic rules may be preferable to avoid unnecessary risk. In both cases, the principle applies, yet the resulting actions differ. This illustrates a key challenge: interpreting abstract ethical principles in real-world scenarios is nuanced, context-dependent, and often controversial, making fully automated ethical planning difficult. We aim to develop an interactive software platform, based on existing work, that encourages human-machine collaboration to interpret these principles in a given classical planning problem and generate plans that not only achieve goals, but also consider the ethics of the plan that achieves such goals.

Computational Machine Ethics (CME) approaches are often divided into top-down, bottom-up, and hybrid approaches (Zhong et al. 2025). Top-down methods (Vanderelst and Winfield 2018; Pagnucco et al. 2021; Grandi et al. 2023) specify rules or guidelines in advance, ensuring transparency but lacking adaptability. Bottom-up approaches (Jiang et al. 2025; Li, Cai, and Xiao 2025) rely on data to infer ethical behaviour, trading off interpretability for flexibility. Hybrid approaches (Allen, Smit, and Wallach 2005; Ramanayake and Nallur 2024) attempt to combine these strengths, but typically still require extensive manual effort to encode ethical rules or examples. Advances in Large Language Models (LLMs) offer a practical means to reduce the manual effort of encoding such rules or examples, which we consider in a planning context.

Recent work has explored incorporating LLMs into automated planning in various ways (Pallagani et al. 2024). Beyond attempts to use LLMs to generate plans directly, they have been applied to facilitate planning processes, including model construction (Oswald et al. 2024), human-LLM collaboration (Wu, Ai, and Hsu 2023), and translation of natural language into structured languages (Ichter et al. 2022; Liu et al. 2023; Favier et al. 2025; Zhong, Song, and Pagnucco 2025). Few contemporary studies leverage LLMs to support automated planning with explicit specifications (Favier et al. 2025; Zhong, Song, and Pagnucco 2025). Favier et al. (2025) use LLMs to decompose and encode general natural language constraints in PDDL3, while Zhong, Song, and Pagnucco (2025) translate high-level ethical principles into context-specific rules represented as action costs in PDDL. Although the latter targets ethics—an underexplored area in automated planning—it lacks a user-facing interface, which Favier et al. (2025) provides. We present *Principles2Plan*, a prototype that enables users to generate ethical plans. While prior work lies at the intersection of users, LLMs, and automated planning, no existing system supports collaborative human-LLM refinement and operationalisation of ethical principles. Principles2Plan addresses this gap by integrating an interactive interface with the pipeline introduced by Zhong, Song, and Pagnucco (2025).

Principles2Plan is a prototype that leverages LLMs and human oversight to incorporate ethical considerations into automated planning. Building on the human-in-the-loop pipeline introduced by Zhong, Song, and Pagnucco (2025), the system takes user input, which an LLM uses to generate context-specific ethical rules from high-level principles. Users can then refine and prioritise these rules before they are encoded and supplied to a classical planner. This design

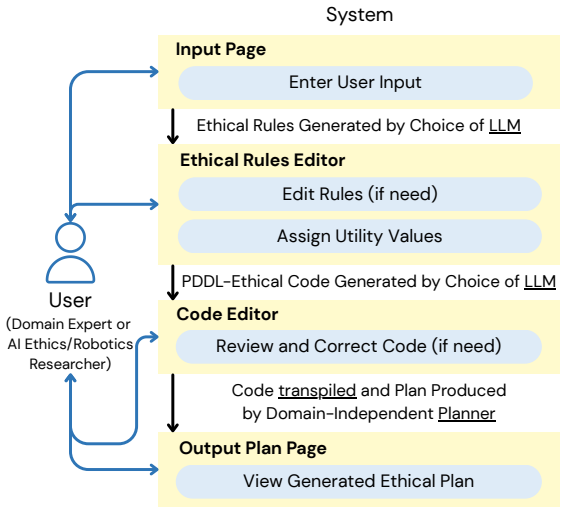

Figure 1: Overall user/system flow.

makes explicit how abstract principles are operationalised into actionable rules to guide planning, enabling transparent and ethically informed plans in real time. By emphasising interactivity and usability, Principles2Plan contributes a system demonstration of how LLMs can bridge the gap between high-level principles and lower-level automated planning, showcasing a novel research direction in CME.

System Overview

To generate an ethical plan from a planning problem and high-level ethical principles, Principles2Plan guides users through four steps on dedicated pages: providing input, reviewing and prioritising generated rules, and reviewing code before producing an ethically-informed plan. Figure 1 illustrates this process from the user's perspective, which we describe in detail in this section. The intended users of the system are domain experts in ethically-sensitive domains, AI ethics and robotics researchers, and anyone interested in the intersection of ethics, LLMs, and automated planning. As the process includes reviewing code and generating plans, users are assumed to have a basic understanding of automated planning and familiarity with PDDL[1] (a standardised language used in planning) and PDDL-Ethical (an extension for ethical constructs) (Jedwabny 2022). We recognise that intended users may lack planning and PDDL expertise; reducing this need remains a challenge for future work.

Input Page The input page of Principles2Plan lets users start generating ethically-informed plans by providing key problem information. These inputs drive the system to generate context-specific ethical rules in natural language, following a structure defined in (Zhong, Song, and Pagnucco 2025). Each rule includes *ethical features*, representing positive or negative ethical characteristics of the rule (e.g., dishonesty as a negative feature). Users can upload and preview their *problem.pddl* and *domain.pddl* files. The user also specifies the initial state, assumptions about the problem or domain, and high-level ethical principles to guide rule generation. Finally, the user can select a preferred model. The system then processes all inputs and prompts the LLM to produce relevant ethical rules in real time.

To help users explore and experiment with the system, Principles2Plan provides multiple example problems across three ethically-sensitive domains: autonomous vehicles, elderly care, and firefighting/rescue. Users can select these examples to populate the input fields directly.

Ethical Rules Editor Since ethical rules generated by an LLM may be inconsistent or imperfect, the next step allows users to review and refine them. Users can add missing rules, remove inappropriate ones, and modify existing rules. To support this process, the system provides explanations from the LLM, detailing the reasoning behind why each rule was generated based on the problem and the specified ethical principle(s). Once users are satisfied with the rules, they can prioritise them by assigning a significance level (1–5) to each ethical feature associated with a rule. The system highlights positive and negative features, allowing users to click on and adjust their importance easily. These rules are then fed into the LLM to generate PDDL-Ethical code, which users review on the following page.

Code Editor On the code editing page, users review the syntax-highlighted PDDL-Ethical code generated from the natural language ethical rules. The code is then transpiled (using the method from (Jedwabny 2022)) into raw PDDL with action costs and submitted to a domain-independent classical planner (Fast Downward). A view of the ethical rules from the previous page is provided alongside to support cross-checking, helping users ensure correctness and consistency between the rules and the code.

Output Plan Page The plan generated with ethical rules and another produced by the same planner using the original problem and domain files are displayed side-by-side, allowing users to directly evaluate the impact of the ethical rules.

One may question the practicality and performance of LLM-generated outputs here and whether they add more work for the user. The performance of the method has been evaluated with DeepSeek-R1-Distill-Llama-70B in (Zhong, Song, and Pagnucco 2025) using metrics including Sentence-BERT similarity (0.82) for generated rules and code generation success rate (82.2%). While these results are not exceptional, they indicate a promising direction. As this is the first implemented prototype of its kind, it may require more human intervention in its current form. We are optimistic that future iterations will improve the balance of collaboration between humans and LLMs.

Conclusion

Principles2Plan is a novel prototype that enables ethically-aware automated planning by combining human guidance with LLMs. Users can generate, refine, and prioritise context-specific ethical rules to produce transparent and ethically-informed plans in real time. Future work will enhance human-LLM collaboration through iterative dialogue and suggestions. Overall, Principles2Plan serves as a hands-on platform for generating ethical plans and for researchers to experiment with interactive ethical decision-making.

[1]https://planning.wiki/guide/whatis/pddl

References

Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7: 149–155.

Favier, A.; La, N.; Verma, P.; and Shah, J. 2025. A Collaborative Numeric Task Planning Framework based on Constraint Translations using LLMs. In *ICAPS 2025 Workshop on Human-Aware and Explainable Planning*.

Grandi, U.; Lorini, E.; Parker, T.; and Alami, R. 2023. Logic-Based Ethical Planning. In *AIxIA 2022—Advances in Artificial Intelligence*, 198–211.

Ichter, B.; Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; Kalashnikov, D.; Levine, S.; Lu, Y.; Parada, C.; Rao, K.; Sermanet, P.; Toshev, A. T.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Yan, M.; Brown, N.; Ahn, M.; Cortes, O.; Sievers, N.; Tan, C.; Xu, S.; Reyes, D.; Rettinghouse, J.; Quiambao, J.; Pastor, P.; Luu, L.; Lee, K.-H.; Kuang, Y.; Jesmonth, S.; Jeffrey, K.; Ruano, R. J.; Hsu, J.; Gopalakrishnan, K.; David, B.; Zeng, A.; and Fu, C. K. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *6th Annual Conference on Robot Learning*.

Jedwabny, M. 2022. *A preference-based approach to machine ethics for automated planning*. Ph.D. Thesis., Universite de Montpellier.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J. T.; Levine, S.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Hessel, J.; Borchardt, J.; Sorensen, T.; Gabriel, S.; Tsvetkov, Y.; Etzioni, O.; Sap, M.; Rini, R.; and Choi, Y. 2025. Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 7(1): 145–160.

Li, J.; Cai, M.; and Xiao, S. 2025. Reinforcement learning-based motion planning in partially observable environments under ethical constraints. *AI and Ethics*, 5(2): 1047–1067.

Liu, B.; Jiang, Y.; Zhang, X.; Liu, Q.; Zhang, S.; Biswas, J.; and Stone, P. 2023. LLM+P: Empowering Large Language Models with Optimal Planning Proficiency. *arXiv preprint arXiv:2304.11477*.

Oswald, J.; Srinivas, K.; Kokel, H.; Lee, J.; Katz, M.; and Sohrabi, S. 2024. Large language models as planning domain generators. In *Proceedings of the Thirty-Fourth International Conference on Automated Planning and Scheduling*, volume 34, 423–431.

Pagnucco, M.; Rajaratnam, D.; Limarga, R.; Nayak, A.; and Song, Y. 2021. Epistemic Reasoning for Machine Ethics with Situation Calculus. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 814–821.

Pallagani, V.; Muppasani, B. C.; Roy, K.; Fabiano, F.; Loreggia, A.; Murugesan, K.; Srivastava, B.; Rossi, F.; Horesh, L.; and Sheth, A. 2024. On the prospects of incorporating large language models (LLMs) in automated planning and scheduling (APS). In *Proceedings of the Thirty-Fourth International Conference on Automated Planning and Scheduling*, volume 34, 432–444.

Ramanayake, R.; and Nallur, V. 2024. Implementing Prosocial Rule Bending in an Elder-Care Robot Environment. In *Social Robotics*, 230–239. Springer Nature.

Vanderelst, D.; and Winfield, A. F. T. 2018. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48: 56–66.

Wu, Z.; Ai, B.; and Hsu, D. 2023. Integrating Common Sense and Planning with Large Language Models for Room Tidying. In *RSS 2023 Workshop on Learning for Task and Motion Planning*.

Zhong, T.; Song, Y.; Limarga, R.; and Pagnucco, M. 2025. Computational Machine Ethics: A Survey. *Journal of Artificial Intelligence Research*, 82: 1581–1628.

Zhong, T.; Song, Y.; and Pagnucco, M. 2025. Generation of Ethical Rules Using Large Language Models. In *AI 2025: Advances in Artificial Intelligence*, 67–79.