Bad AI, Good AI: Rethinking the Agency of Our Artificial Teammates

Reuth Mirsky

Computer Science Department
Tufts University
reuth.mirsky@tufts.edu

A prevalent assumption in human-robot and human-AI teaming is that artificial teammates should be compliant and obedient. In this talk, I will question this assumption by presenting the Guide Robot Grand Challenge (Mirsky and Stone 2021) and discussing the components required to design and build a service robot that can intelligently disobey. This challenge encompasses a variety of research problems, as I will exemplify via three challenges: reasoning about the goals of other agents, choosing when to interrupt, and interacting in a tightly coupled physical environment.

For the first challenge, reasoning about the goals of other agents, I will present our work on **goal recognition as reinforcement learning** (Amado, Mirsky, and Meneguzzi 2022): Most approaches for goal recognition rely on specifications of the possible dynamics of the actor in the environment when pursuing a goal. These specifications suffer from several issues. First, encoding these dynamics requires careful design by a domain expert, which is often not robust to noise at inference time. Second, existing approaches often need costly real-time computations to determine the likelihood of each goal. Here, I will present our framework that combines model-free reinforcement learning and goal recognition to alleviate the need for careful, manual domain design and the need for costly online executions. This framework consists of two stages: offline learning of policies or utility functions for each potential goal and online inference.

For the second challenge, choosing when to interrupt, I will present a recent contribution exploring **interruptions in human-robot teaming** (Mannem et al. 2023): Productive and efficient human-robot teaming is a highly desirable ability in service robots, yet a robot needs to consider a fundamental trade-off in such tasks. On the one hand, gaining information from communication with teammates can help individual planning. On the other hand, such communication comes at the cost of distracting teammates from efficiently completing their goals, which can also harm the overall team performance. In this study, we quantify the cost of interruptions in terms of degradation of human task performance, as a robot interrupts its teammate to gain information about their task. Interruptions are varied in timing, content, and proximity. The results show that people find the interrupting robot significantly less helpful even when there was no objective decline in team performance. These research outcomes can inform numerous applications where collaborative robots must be aware of the costs and gains of interruptive communication, including logistics and service robots.

The third challenge tackles the challenge of a disobedient robot **acting in a tightly-coupled physical environment (Ghonasgi et al. 2022)** and needs to determine the abilities of its handler. In this project, we draw upon existing literature on human skill assessment and present extrinsic and intrinsic performance metrics that quantify how the human-exoskeleton system's behavior changes over time. Specifically, we present new performance metrics and new evaluation domains to provide insight into the shared system's kinematics associated with 'successful' movements. Changes in the newly developed kinematics-based measure further illuminate how the participant's intrinsic behavior is altered over the training period. Thus, we can quantify the changes in the human-exoskeleton system's behavior observed in relation with learning.

Finally, I will briefly discuss the many remaining challenges to achieving intelligent disobedience in AI and how I plan to tackle them in my research.

References

Amado, L.; Mirsky, R.; and Meneguzzi, F. 2022. Goal recognition as reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9644–9651.

Ghonasgi, K.; Mirsky, R.; Haith, A. M.; Stone, P.; and Deshpande, A. D. 2022. Quantifying Changes in Kinematic Behavior of a Human-Exoskeleton Interactive System. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10734–10739. IEEE.

Mannem, S.; Macke, W.; Stone, P.; and Mirsky, R. 2023. Exploring the Cost of Interruptions in Human-Robot Teaming. In *The 2023 IEEE-RAS International Conference on Humanoid Robots (Humanoids)*.

Mirsky, R.; and Stone, P. 2021. The seeing-eye robot grand challenge: rethinking automated care. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*.