Building Efficient, Reliable, and Ethical Autonomous Systems

Justin Svegliato

The goal of my research is to build autonomous systems that can operate in natural, partially observable environments for long periods of time in an *efficient*, *reliable*, and *ethical* way. At a high level, my research focuses on three important areas of autonomous systems: efficient planning, reliable execution, and ethical compliance. To explore these areas, I develop approaches based on MDPs, POMDPs, and Dec-POMDPs along with their corresponding solution methods by using dynamic programming, linear programming, reinforcement learning, state abstractions, heuristic search, and deep learning. The impact of my research has fortunately been underscored by a **distinguished paper award** (AAAI), **strong publications** (AAAI/IJCAI/ECAI/IROS/ICRA/AIES/AAMAS), an **ongoing industry collaboration** (Nissan Research Center), **mentorship opportunities** (BS/MS/PhD students), an **NSF Graduate Research Fellowship**, and an **NSF Grant on Adaptive Metareasoning for Bounded Rational Agents**.

Metareasoning for efficient planning and reliable execution It has long been recognized that autonomous systems cannot have perfect rationality due to the intractability of optimal decision making in complex domains [1]. In response, there have been substantial efforts to develop computational approaches to bounded rationality. *Metareasoning*, a particularly effective computational approach to bounded rationality, enables an autonomous system to optimize—specifically monitor and control—its own planning and execution processes in order to act effectively in its environment [2]. This allows the autonomous system to handle any uncertainty about the range of its circumstances and the limitations of its capabilities. Consequently, due to the growth in the autonomy, complexity, and generality of artificial intelligence and robotics over the years, metareasoning has become critical to autonomous systems.

My dissertation proposes a metareasoning framework for **efficient planning** and **reliable execution** in autonomous systems. This framework enables an autonomous system to optimize its planning processes that compute a policy and its execution processes that follow a policy. In particular, by monitoring and controlling its own planning and execution processes, the autonomous system not only *efficiently computes a policy* by, say, generating the highest quality policy under strict time constraints but also *reliably follows that policy* by, say, recovering from unanticipated scenarios and addressing safety concerns. For example, a self-driving car must initially compute a route plan by balancing route time with computation time and later follow that route plan by recovering from unanticipated scenarios that impede its path and addressing safety concerns that endanger its passengers [3]. My research on efficient planning and reliable execution has led to work on optimal stopping for anytime planning [4, 5, 6], optimal hyperparameter tuning for anytime planning [7, 8], partial state abstractions [9], exception recovery [10, 11], and safe operation [12].

Models and algorithms for ethical compliance Most importantly, while my dissertation focuses on metareasoning for efficient planning and reliable execution in autonomous systems, my research recognizes that artificial intelligence and robotics have rapidly been deployed in sociocultural environments that often have a serious impact on society [13]. Generally, I propose models and algorithms for **ethical compliance** that enable autonomous systems to align with the values of their stakeholders. In particular, an *ethically compliant autonomous system*—a framework with a decision-making model designed for a task along with an ethical context and a moral principle designed for an ethical theory—*optimizes completing a task while following an ethical theory*. For instance, an elder care robot must complete a medical diagnostic task while following an ethical theory, such as Kantianism, utilitarianism, or virtue ethics, to tailor its support based on the state of the patient so as to reduce the risk of injury or the loss of dignity [14]. My research on ethical compliance has led to a body of work on ethically compliant autonomous systems [15, 16, 17, 18].

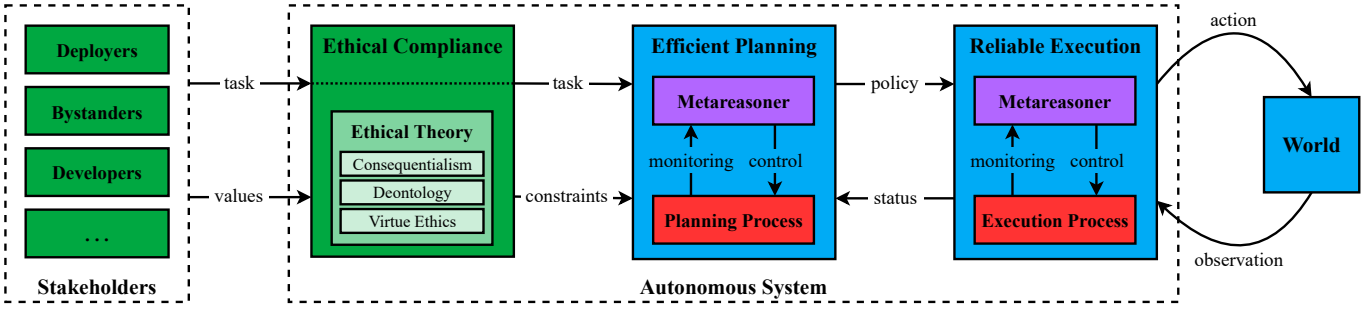

Figure 1: An architecture for autonomous systems with ethical compliance, efficient planning, and reliable execution.

Figure 1 proposes an architecture for an autonomous systems with three distinct modules for ethical compliance, efficient planning, and reliable execution. First, the **ethical compliance** module *builds* a set of ethical constraints influenced by an ethical theory that attempts to align with the values of the stakeholders and then *interacts* with the efficient planning module by sending the task of the stakeholders and the set of ethical constraints that will constrain the behavior of the autonomous system. Next, the **efficient planning** module *runs* a metareasoner that monitors and controls the planning process to efficiently compute a policy and then *interacts* with the reliable execution module by sending the policy and receiving a status that may trigger recomputing a new policy for the autonomous system. Finally, the **reliable execution** module *runs* a metareasoner that monitors and controls the execution process to reliably follow a policy and then *interacts* with the world by performing actions and making observations.

Selected Past, Present, and Future Research Projects

My research focuses on a range of research projects that are important to building autonomous systems with efficient planning, reliable execution, and ethical compliance. This includes work that can be completed by undergraduate and graduate students from different academic backgrounds (computer science, mechanical engineering, electrical engineering, and psychology) and work that encourages an ongoing industry collaboration that I have developed over the years (Nissan Research Center). I describe selected research projects at various stages of development below.

Implementing ethical theories within autonomous systems [*Ethical Compliance*] In recent work, we have developed a formal approach to building autonomous systems that comply with an ethical theory. In general, a naive approach to enabling an autonomous system to comply with an ethical theory is to modify the objective function of its decision-making model. Modifying this objective function, however, may cause the autonomous system to fail to reflect the values of its stakeholders in two significant ways [19, 20, 21]. First, adjusting the objective function can lead to unpredictable effects on the behavior of the autonomous system due to the complexity of its decision-making model. Second, using the objective function to represent both a task and an ethical theory can result in incommensurable conversions as it blends them within the decision-making model implicitly. As a result, we have proposed an *ethically compliant autonomous system* that optimizes completing a task while following an ethical theory [15, 16, 17, 18]. This system decouples task completion from ethical compliance by describing the task as a decision-making model and the ethical theory as an ethical context and a moral principle. This is formally expressed as a mathematical program with primary constraints that encode the task and secondary constraints that encode the ethical theory. As a demonstration, we showcased its effectiveness on an autonomous driving domain across a range of ethical theories for divine command theory, prima facie duties, and virtue ethics. **Future Research:** We plan to apply our approach to more complex ethical theories, such as Kantianism, utilitarianism, social contract theory, and natural law theory, on more complex domains not only in simulation but also on an actual mobile robot.

Determining the optimal stopping point of anytime planning [*Efficient Planning*] Early on and throughout my research, we have developed several decision-theoretic metareasoning techniques for determining when to interrupt an anytime planner and act on its current plan. At a high level, autonomous systems use anytime planners that offer a trade-off between solution quality and computation time that has proven to be useful for real-time planning. To optimize this trade-off, an autonomous system must determine when to interrupt the anytime planner and act on its current plan. Existing techniques for determining the stopping point of an anytime planner rely on planning with a performance profile that must be compiled offline [22]. Planning with a performance profile, however, imposes many assumptions often violated by autonomous systems in complex domains. Hence, we have introduced two techniques for estimating the optimal stopping point of an anytime planner that can be used under two different conditions. Intuitively, when the performance characteristics of the anytime planner are known, the first technique estimates the optimal stopping point online by predicting its future performance based on its past performance using online performance prediction [4]. However, when the performance characteristics of the anytime planner are unknown, the second technique estimates the optimal stopping point online by learning its true performance using reinforcement learning [5, 6]. **Future Research:** We plan to explore more informed techniques for online performance prediction and more efficient reinforcement learning methods in mobile robot task, path, and motion planning domains.

Tuning the hyperparameters of anytime planning [*Efficient Planning*] Extending our work on optimal stopping for anytime planning, we have started to develop a decision-theoretic metareasoning technique for adjusting the hyperparameters of an anytime planner at runtime. While there are techniques for adjusting the hyperparameters of

specific anytime planners at runtime [23, 24], they require expertise in the anytime planner and also lack generality and formal analysis. Thus, we have been developing a decision-theoretic metareasoning technique for optimal hyperparameter tuning of an anytime planner online [7, 8]. This techniques expresses the metareasoning problem as a deep reinforcement learning problem with two main attributes: *states* that reflect the quality and computation time of the current solution along with any other features needed to summarize the internal state of the anytime planner, the instance of the problem, or the performance of the underlying system and *actions* that reflect tuning the internal hyperparameters of the anytime planner. **Future Research:** We plan to apply our approach to an anytime task planner based on A* and an anytime motion planner based on RRT* across a range of common mobile robot domains.

Solving large MDPs with partial state abstractions [*Efficient Planning*] To reduce the complexity of the models solved by anytime planners, we have developed a formal algorithm for solving large MDPs. Given the need to use many state features in MDPs for autonomous systems to behave effectively in complex domains, MDPs must often be solved approximately in real-time settings given the exponential growth of the state space in the number of state factors. However, while there are techniques that use state abstractions in MDPs to reduce the complexity of the state space, they often eliminate details that are required to produce effective behavior in autonomous systems. Consequently, we have offered a solution method for solving a large *ground MDP* that performs two phases [9]: it initially (1) sketches a policy using an *abstract MDP* and later (2) refines that policy using different *partially abstract MDPs* that each compress ground states to condense irrelevant details but expand abstract states to retain relevant details. As a demonstration, we highlighted its scalability by computing near-optimal policies to large MDPs in minutes rather than hours in an earth observation satellite domain. **Future Research:** We plan to develop decision-theoretic metareasoning techniques for automatically determining the abstracts states to be expanded and the ground states to be compressed in mobile robot task and path planning domains.

Recovering from exceptions [*Reliable Execution*] After shifting from metareasoning for planning to execution, we have developed a belief space metareasoning approach to building autonomous systems that can detect, identify, and handle exceptions during operation. Resolving exceptions that violate the assumptions of the decision-making model of an autonomous system is challenging for three reasons. First, because an exceptional scenario is not captured by definition, the decision-making model does not have the information needed to resolve that exception. Second, while the decision-making model can be extended to capture an exceptional scenario, this will rapidly grow the complexity of the decision-making model for each exception. Third, since a decision-making model cannot capture every exceptional scenario, there will always be exceptions that cannot be resolved properly. Although early work on exception recovery has focused on detecting and identifying exceptions [25, 26], they do not offer a framework that can also handle exceptions without human assistance. Therefore, we have proposed an *exception recovery metareasoning system* that interleaves a main decision process designed for normal operation with many exception handlers designed for exceptional operation using a belief over exceptions [10, 11]. As a demonstration, we showed that—with each new exception handler—our approach decreases its reliance on human assistance while increasing its utility in an autonomous driving domain both in simulation and on an autonomous vehicle prototype. **Future Research:** We plan to develop more complex exception handlers with formal analysis that can resolve multiple simultaneous exceptions.

Maintaining and restoring safe operation [*Reliable Execution*] By building on our work on exception recovery, we have started to develop a metareasoning approach for optimizing safety in autonomous systems. While planning and robotics experts carefully design, build, and test the models used by autonomous systems for high-level decision making, it is infeasible for these models to ensure safety across every scenario within the domain of operation [27]. A naive approach to maintaining and restoring safety is to use an exhaustive decision-making model with every feature needed to cover every scenario within the domain of operation [28]. This comprehensive model, however, is infeasible since it would not only be impossible to build in complex domains but also impossible to solve with exact or even approximate solution methods in real-time settings. Accordingly, we have been developing an approach to building a *safety metareasoning system* that mitigates the *severity* of the system's safety concerns while reducing the *interference* to the system's task: the system executes a *task process* and a set of *safety processes* in parallel, where the task process completes the task while the safety processes each address a safety concern, arbitrating with a conflict resolver [12]. As a demonstration, we showed that the system mitigates the severity of safety concerns while reducing interference to the task in a planetary rover exploration domain. **Future Research:** We plan to develop formal analysis and apply our approach to an autonomous space station domain to show its effectiveness in simulation.

References

[1] S. J. Russell and E. H. Wefald, *Do the Right thing: Studies in Limited Rationality*. Cambridge, MA: MIT Press, 1991.

[2] S. Zilberstein, "Metareasoning and bounded rationality," in *Metareasoning: Thinking about Thinking*, Cambridge, MA: MIT Press, 2011.

[3] C. Basich, J. Svegliato, K. H. Wray, S. Witwicki, J. Biswas, and S. Zilberstein, "Learning to optimize autonomy in competence-aware systems," in *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[4] J. Svegliato, K. H. Wray, and S. Zilberstein, "Meta-level control of anytime algorithms with online performance prediction," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[5] J. Svegliato and S. Zilberstein, "Adaptive metareasoning for bounded rational agents," in *Proceedings of the IJCAI Workshop on Architectures and Evaluation for Generality, Autonomy and Progress in AI (AEGAP)*, 2018.

[6] J. Svegliato, P. Sharma, and S. Zilberstein, "A model-free approach to meta-level control of anytime algorithms," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[7] A. Bhatia, J. Svegliato, and S. Zilberstein, "On the benefits of randomly adjusting anytime weighted A*," in *Proceedings of the 14th Symposium on Combinatorial Search (SoCS)*, 2021.

[8] A. Bhatia, J. Svegliato, and S. Zilberstein, "Metareasoning for adjustable algorithms with deep reinforcement learning," in *Submission to the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[9] S. Nashed B*, J. Svegliato*, M. Brucato, C. Basich, R. Grupen, and S. Zilberstein, "Solving Markov decision processes with partial state abstractions," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[10] J. Svegliato, K. H. Wray, S. J. Witwicki, J. Biswas, and S. Zilberstein, "Belief space metareasoning for exception recovery," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[11] J. Svegliato, S. J. Witwicki, K. H. Wray, and S. Zilberstein, "Introspective autonomous vehicle operational management," U.S. Patent 10,649,453, May 2020.

[12] J. Svegliato, C. Basich, S. Saisubramanian, and S. Zilberstein, "Metareasoning for optimizing safety in autonomous systems," in *Submission to the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[13] V. Charisi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A. F. Winfield, and R. Yampolskiy, "Towards moral autonomous systems," in *arXiv preprint arXiv:1703.04741*, 2017.

[14] J. Shim, R. Arkin, and M. Pettinatti, "An intervening ethical governor for a robot mediator in patient-caregiver relationship," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[15] J. Svegliato, S. B. Nashed, and S. Zilberstein, "An integrated approach to moral autonomous systems," in *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020.

[16] J. Svegliato, S. B. Nashed, and S. Zilberstein, "Ethically compliant planning in moral autonomous systems," in *Proceedings of the IJCAI Workshop on AI Safety (AISafety)*, 2020.

[17] J. Svegliato, S. B. Nashed, and S. Zilberstein, "Ethically compliant sequential decision making," in *Proceedings of the 35th AAAI International Conference on Artificial Intelligence (AAAI)*, 2021.

[18] S. Nashed, J. Svegliato, and S. Zilberstein, "Ethically compliant planning within moral communities," in *Proceedings of the 4th Conference on AI, Ethics, and Society (AIES)*, 2021.

[19] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Wiley Online Library, 2016.

[20] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, "Alignment for advanced machine learning systems," in *Machine Intelligence Research Institute*, 2016.

[21] D. Hadfield-Menell and G. K. Hadfield, "Incomplete contracting and AI alignment," in *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2019.

[22] E. A. Hansen and S. Zilberstein, "Monitoring and control of anytime algorithms: A dynamic programming approach," *Artificial Intelligence (AIJ)*, 2001.

[23] E. A. Hansen and R. Zhou, "Anytime heuristic search," *Journal of Artificial Intelligence Research (JAIR)*, 2007.

[24] J. Thayer and W. Ruml, "Using distance estimates in heuristic search," in *Proceedings of the 19th International Conference on Automated Planning and Scheduling (ICAPS)*, 2009.

[25] P. Goel, G. Dedeoglu, S. I. Roumeliotis, and G. S. Sukhatme, "Fault detection and identification in a mobile robot using multiple model estimation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2000.

[26] V. Verma, G. Gordon, R. Simmons, and S. Thrun, "Particle filters for fault diagnosis," *IEEE Robotics and Automation Magazine*, 2004.

[27] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv:1606.06565*, 2016.

[28] S. Saisubramanian, E. Kamar, and S. Zilberstein, "A multi-objective approach to mitigate negative side effects," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.