Building Efficient, Reliable, and Ethical Autonomous Systems

Justin Svegliato

As autonomous systems rapidly grow in adaptability, effectiveness, and sophistication, their deployment has been accelerating in complex real-world domains that range from elder care robotics and autonomous driving to smart city management and military technology. However, while our ability to build *efficient* and *reliable* autonomous systems that integrate into our daily lives has expanded over the years, it has inevitably outstripped our ability to build *ethical* autonomous systems. Therefore, the goal of my research is to build autonomous systems that operate in natural, partially observable, stochastic domains for long durations in not only an *efficient* and *reliable* but also *ethical* way. Given the goal of my research, I have developed a range of approaches based on MDPs, POMDPs, and Dec-POMDPs along with their solution methods by using dynamic programming, mathematical programming, reinforcement learning, machine learning, deep learning, abstractions, heuristic search, and probabilistic graphical models. As a demonstration of my research, I have applied these approaches to autonomous vehicles (route navigation, obstacle handling, lane merging, and intersection negotiation), planetary exploration rovers, earth observation satellites, and standard general-purpose mobile robots.

My research has fortunately led to a distinguished paper award (AAAI), an NSF Graduate Research Fellowship, and top publications (AAAI/IJCAI/ECAI/ICRA/IROS/AAMAS/AIES/SoCS). Moreover, I have had the opportunity to mentor students, write a grant proposal, and collaborate with industry. First, I have been a mentor for 6 BS, MS, and PhD students who have all had the experience as an author on multiple papers for top conferences. In fact, most recently, I am excited to say that one of my BS students will start a PhD in artificial intelligence at Brown University in the fall. Next, I was a main author on an NSF grant proposal on adaptive metareasoning for bounded rational agents that was awarded over $400,000. Finally, I am a key contributor to an industry collaboration with Nissan Research Center that has led to top publications and patents for over 3 years.

Research Directions

To advance my goal of building efficient, reliable, and ethical autonomous systems, my main research directions work toward (1) bounded rationality through metareasoning for *efficient planning* and *reliable execution* as well as (2) value alignment through models and algorithms for *ethical compliance*. I outline my main research directions and also how they integrate with each other below.

Toward bounded rationality through metareasoning for efficient planning and reliable execution

It has long been recognized that autonomous systems cannot have *perfect rationality* due to the computational 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 to operate more effectively in its environment [2]. This enables the autonomous system to manage any uncertainty about the range of its circumstances and the limitations of its capabilities. As a result, given the complexity inherent to natural, partially observable, stochastic domains, metareasoning as a computational approach to bounded rationality has become critical to autonomous systems.

Recently, in my dissertation, I propose 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 available 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 interest in metareasoning for efficient planning and reliable execution has led to my 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, 13].

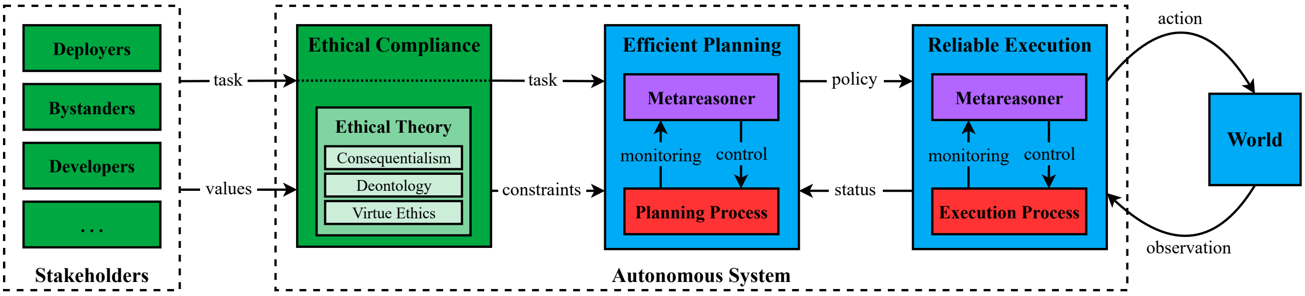

Figure 1: An architecture for autonomous systems that illustrates the connections between each area of my research.

Toward value alignment through models and algorithms for ethical compliance

Most importantly, while my dissertation focuses on metareasoning for efficient planning and reliable execution, my research recognizes that autonomous systems have rapidly been deployed in sociocultural domains that have a serious impact on society [14]. Generally, I offer models and algorithms for ethical compliance that enable autonomous systems to align with the values of their stakeholders. In particular, I propose *ethically compliant autonomous systems* that optimize *completing a task* subject to *following an ethical theory* by decoupling a decision-making model that describes the task from an ethical context and a moral principle that describe the ethical theory. For instance, an elder care robot must complete a medical task while following an ethical theory, such as Kantianism, utilitarianism, or virtue ethics, to tailor its support based on the physical or mental state of the patient so as to reduce the risk of injury or the loss of dignity [15]. My interest in models and algorithms for ethical compliance has led to my body of work on ethically compliant autonomous systems [16, 17, 18, 19].

Toward integrating bounded rationality with value alignment

Figure 1 proposes an architecture for autonomous systems with three distinct modules for ethical compliance, efficient planning, and reliable execution in order to demonstrate how each area of my research interacts with each other. First, the ethical compliance module (1) *builds* a set of ethical constraints from a given ethical theory that attempts to align with the specified values and then (2) *interacts* with the efficient planning module by sending the specified task and the set of ethical constraints that will constrain the policy of the autonomous system. Next, the efficient planning module (3) *runs* a metareasoner that monitors and controls the planning process to efficiently compute a policy and then (4) *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 (5) *runs* a metareasoner that monitors and controls the execution process to reliably follow a policy and then (6) *interacts* with the world by performing actions and making observations.

Research Results

My main research results have focused on my goal of building efficient, reliable, and ethical autonomous systems. This includes research projects that can be completed by undergraduate and graduate students from various academic backgrounds in philosophy, mechanical engineering, electrical engineering, sociology, psychology, and computer science. I describe my main research results that span across *efficient planning*, *reliable execution*, and *ethical compliance* and also summarize their potential future work below.

Embedding ethical theories into autonomous systems *Ethical Compliance*

Concerned with the ethical implications of artificial intelligence, we have developed a novel approach to building autonomous systems that can comply with an ethical theory that we have shown to be more accurate and practical than existing techniques. In general, a simple approach to enabling an autonomous system to comply with an ethical theory is to modify the objective function of its decision-making model directly. Modifying this objective function, however, may cause the autonomous system to fail to reflect the values of its stakeholders in two main ways [20, 21, 22]. 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 developed an *ethically compliant autonomous system* that optimizes completing a task subject to following an ethical theory [16, 17, 18, 19]. 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. On an autonomous driving domain across a range of ethical theories for divine command theory, prima facie duties, and virtue ethics, we show that our approach is more accurate and practical than existing techniques in both simulation and a user study of experts. **Future Work:** We plan to expand our approach to more complex ethical theories, such as Kantianism, utilitarianism, and natural law theory, and more challenging domains on an actual mobile robot.

Determining the optimal stopping point of anytime planners online *Efficient Planning*

Inspired by early work on metareasoning for anytime planning, we have developed novel decision-theoretic metareasoning for determining the optimal stopping point of an anytime planner online that outperforms existing techniques that rely on extensive offline work. At a high level, autonomous systems often use anytime planners that offer a trade-off between solution quality and computation time that has proven to be useful for real-time planning problems. 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 optimal stopping point of an anytime planner have typically relied on planning with a performance profile offline [23]. Planning with a performance profile, however, imposes many assumptions that are often violated by autonomous systems in complex domains because it can involve many hours or even days of extensive offline work. Hence, we developed two metareasoning techniques that can be used under two different conditions to estimate the optimal stopping point of an anytime planner online. 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 with 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 with reinforcement learning [5, 6]. Using common anytime planners across standard real-time planning problems, we show that our approach outperforms existing techniques that rely on extensive offline work, which reduces the overhead and increases the benefits of using anytime planning in autonomous systems. **Future Work:** We plan to build more informed online performance prediction methods and more efficient reinforcement learning methods for common mobile robot task, path, and motion planning domains.

Adjusting the hyperparameters of anytime planners online *Efficient Planning*

Extending our recent work on optimal stopping for anytime planning, we have developed novel decision-theoretic metareasoning for adjusting the hyperparameters of an anytime planner online that boosts the performance of anytime planning and eliminates any need for manual hyperparameter tuning. Generally, while there are many methods for adjusting the hyperparameters of *specific* anytime planners online [24, 25], they require expertise of the anytime planner and also lack generality or formal analysis. Thus, we developed a metareasoning technique that learns how to adjust the hyperparameter of an anytime planner online by using deep reinforcement learning [7, 8]. Formally, this technique expresses the metareasoning problem as a deep reinforcement learning problem with two main attributes: (1) *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 (2) *actions* that reflect tuning the internal hyperparameters of the anytime planner. Using a common anytime task planner based on A* across standard real-time planning problems, we show that our approach boosts the performance of anytime planning and eliminates any need for manual hyperparameter tuning. **Future Work:** We plan to expand our metareasoning technique to more sophisticated anytime task, path, and motion planners in common mobile robot domains.

Solving large MDPs with partial state abstractions *Efficient Planning*

Reducing the complexity of anytime planning, we have developed a novel algorithm for solving large MDPs with partial state abstractions that calculates near-optimal solutions to real-time planning problems in minutes rather than hours unlike existing techniques that rely on state abstractions. As motivation, 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 due to 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 developed an algorithm 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 and expand abstract states to retain relevant details. On an earth observation satellite domain in simulation, we show that our approach calculates near-optimal solutions to real-time planning problems in minutes rather than hours unlike existing techniques that rely on state abstractions. **Future Work:** We plan to propose decision-theoretic metareasoning for estimating the abstract states to be expanded and the ground states to be compressed in common mobile robot domains.

Recovering from exceptions during operation *Reliable Execution*

Going beyond metareasoning for anytime planning to plan execution, we have developed novel belief-space metareasoning in autonomous systems for not only detecting and identifying but also handling exceptions that outperforms existing techniques relying on human assistance. Resolving exceptions that violate the assumptions of a decision-making model of an autonomous system poses three challenges. First, because an exceptional scenario is not captured by definition, its decision-making model does not have the information needed to resolve that exception. Second, while its decision-making model can be extended to capture an exceptional scenario, this will rapidly grow the complexity of its decision-making model for each exception. Third, since its decision-making model cannot capture every exceptional scenario, there will always be exceptions that cannot be resolved properly. Although work that addresses the challenges of exception recovery has focused on detecting and identifying exceptions [26, 27], they do not offer a framework that can also handle exceptions without human assistance. Therefore, we developed 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]. On an autonomous driving domain both in simulation and on a fully operational autonomous vehicle prototype, we show that our approach decreases its reliance on human assistance and increases its utility while outperforming existing techniques relying on human assistance. **Future Work:** We plan to build more sophisticated exception handlers with provable guarantees on safety that detect, identify, and handle multiple simultaneous exceptions that interact with each other in a complex way.

Maintaining and restoring safety during operation *Reliable Execution*

Building on my recent work in exception recovery, we have developed novel decision-theoretic metareasoning in autonomous systems for maintaining and restoring safety that outperforms existing techniques that rely on a monolithic model that explodes the complexity of the problem. Naturally, 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 [28]. A naive approach to maintaining and restoring safety is to use an exhaustive decision-making model with every feature needed to cover every scenario that can be encountered during operation [29]. Such a comprehensive model, however, is infeasible to use 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 developed 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. That is, the safety metareasoning system executes a *task process* and a set of *safety processes* in parallel such that the task process completes the task while the safety processes each address a safety concern, arbitrating with a conflict resolver [12, 13]. On a planetary rover exploration domain in simulation, we show that our approach optimally mitigates the severity of safety concerns and reduces the interference to the task while outperforming existing techniques that rely on a monolithic model that explodes the complexity of the problem. **Future Work:** We plan to offer a formal analysis of our approach and apply it to an autonomous space station domain potentially in collaboration with NASA.

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] J. Svegliato, C. Basich, S. Saisubramanian, and S. Zilberstein, "Using metareasoning to maintain and restore safety for reliable autonomy," in *Submission to the IJCAI Workshop on Robust and Reliable Autonomy in the Wild (R2AW)*, 2021.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

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

[25] 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.

[26] 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.

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

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

[29] 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.