Research Statement

Leilani H. Gilpin (lgilpin@mit.edu)

My work enables complex machines to use **explanations to dynamically detect and mitigate anomalous behaviors**. This is important as autonomous agents are increasingly deployed in real world settings, where there has been an increase in malfunctions and errors leading to injuries[1] and even deaths[2]. Such level of increased harm on human lives is undesirable and completely untenable. Consequently, assessing, pre-deployment reliability is important for autonomous agents that are responsible for decisions in consequential settings. A key component of pre-deployment verification of the agent is **an explanation**; a model-dependent artifact providing an end-user (technical or otherwise) a reason or justification for the decision of the autonomous agent being assessed. But these explanations are not standard nor do they have a common evaluation metric; they are explaining different processes [4]. Further, after-the-fact explanations are simply one part of how people defend their actions and behaviors; they also help people reason about new scenarios and decisions.

My research utilizes explanations in two distinct ways. System-wide explanations are provided to an end-user for analysis, while internal explanations are used among subsystems to defend their actions and make more robust higher-level decisions. Therefore, I differentiate between an internal subsystem explanation (or internal explanation); the symbolic reasons and dependencies for a specific, local subsystems' behavior, and a system-wide narrative explanation (or system explanation); a (mostly causal) chain of reasoning generated from the underlying subsystems. Since the underlying reasons and dependencies are symbolic, they can be translated into a **human-understandable explanation** at various degrees of detail: for anomaly detection, legal analysis, and diagnosis.

My main idea is that anomalies are not necessarily outliers on graphs, they are circumstances which cannot be explained away by a consensus of neighboring subsystems. Instead, using internal subsystem explanations, anomalies are detected and explained with a higher accuracy. If a subsystem provides an explanation that is inadequate or inappropriate, the subsystem should either corrected or disabled. I apply this methodology to **full-system design** to enable the subsystems of a complex machine to use **introspection and explanation** to make more robust, explainable decisions.

PhD Research - Explaining Opaque Decision Making

The "state of the art" intelligent systems, like Watson, Deep Blue and AlphaZero are opaque to humans. Similar opaque processes are operating dangerous machinery, choosing optimal robotic trajectories, and driving; tasks where small errors can lead to sizable consequences. The key difference between machine and human decision making is that humans can reason about their behavior, provide arguments supporting their actions, and tell a complete, coherent story of what happened and why. In order for us to trust autonomous decision making systems, they will need to possess a similar **ability to testify**: to recount, defend, and explain their behavior.

In my PhD research, I built self-explaining architectures to address the black-box decision-making problem: by **supplementing opaque decisions with commonsense** at various levels of detail. My work is focused on the application of a self-driving vehicle, but my

[1] Mall robot injures a toddler: `https://qz.com/730086/a-robot-mall-cop-did-more-harm-than-good/`

[2] Uber self-driving car pedestrian fatality: `https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html`

architecture ideas could be applied to the design of other complex machines, such as systems security and home robotics. For example, I developed an adaptable monitoring framework called "reasonableness monitors," which can judge and explain whether or not an intended behavior is reasonable in various applications and domains [6]. It uses a commonsense knowledge base and a set of hand-curated rules to construct human-understandable situations. I demonstrated that rules could learned when the monitor explains new situations [5]. I apply the concepts of reasonableness monitoring and rule learning to a full system: a **self-explaining architecture** that models complex systems as a layered system of communicating agents that can **explain their behavior and learn from their mistakes**. Each subsystem is encapsulated in a reasonableness monitor that is consistently explaining why the behavior is reasonable or not. In system-level decisions, the explanations from the underlying parts provide reasons, evidence, and arguments supporting a proposed decision or action. In the next sections, I elaborate on the technical details of reasonableness monitoring, learning from explanations, and applying these methods to a complex machine: a self-explaining full-system architecture for an autonomous vehicle.

Explaining Local Decisions with Reasonableness Monitors

A common failure mode in autonomous agents is a lack of commonsense. Autonomous vehicles can be fooled to proceed through a stop sign with a few stickers [1], and a learning system for soccer can fail when present unexpected scenarios; like a goalkeeper falling down [10]. To address the commonsense problem, I developed a **reasonableness monitor**: a wrapper that encapsulates an opaque process and supplements their intended behaviors with commonsense [6, 7].

Most explanatory systems are domain or system specific, so reasonableness monitors contain a parsing process to represent the inputs in a standard, flexible representation. The input data and rules are parsed into an RDF[3] representation of a combination of primitive actions. For natural language inputs, I use Roger Schanks' Conceptual Dependency Theory [11], which include attributes for the action actor, action object, time, etc. These attributes are represented in the RDF file as symbolic concepts, which are supplemented with previous behavior (history) and commonsense knowledge. In the reasoning step, the monitor checks whether the input RDF violates any constraints, and then produces a human-readable natural language explanation of whether the input is reasonable or not. If is deemed unreasonable, the monitor will continue to query a commonsense knowledge base for alternative context supporting a reasonable state. If no alternative context is appropriate, the input is deemed unreasonable, and an explanation supporting the unreasonable state is provided.

Erroneous and unreasonable cases are not well-represented in training sets, so I built my own data set to evaluate the system. The reasonableness monitor was able to correctly identify and explain both unreasonable and reasonable inputs in both the autonomous driving and image captioning domain. In an user study, participants were satisfied with the explanations, providing an average score of 3.97 out of 5 on a Likert scale from 1 (not convincing) to 5 (very convincing).

[3] `https://www.w3.org/RDF/`

Learning Commonsense Rules from Explanations

Humans have a unique ability to adapt to new contexts, even in a single instance. When driving in new cities or presented with novel scenarios, we are able to identify the situation (state), and explain the most reasonable futures and decisions. We are able to abstract and explain new scenarios with limited amounts of data. It would be difficult and expensive to represent all possible driving decisions in training data. Therefore, it is important to provide driving modules the ability to **abstract and learn from new experiences**.

But providing machines with these kinds of rules are time-consuming to create; they are human-curated and static. So I developed a method to learn behavioral rules from explanations [5] where the rules are abstract and determined from a small set of examples. The method processes a natural language explanation, classifies the context and sentiment of the explanation, and if the explanation represents a new situation without resulting in an error or anomaly, a new rule is created and added to the existing monitoring system. The hypothesis was that learning new rules will improve the existing monitor in new situations and contexts. In the initial evaluation, new rules were able to be learned by querying logs and explanations generated by the monitor around the autonomous planner. In our hardware, MIT RACECAR test, the system was able to correctly differentiate between 5 erroneous, 5 normal, and 5 repeated cases; only creating new rules when cases were repeated.

Explaining System-wide Errors from Vehicle Logs

Autonomous vehicles have been set to ignore certain objects on the road[4] without knowing why. This ended in a fatality in March 2018, when an Uber self-driving test vehicle (a modified 2017 Volvo XC90) struck and killed a pedestrian in Tempe, Arizona. To combat inconsistent information between the parts of a complex machine, I developed a system-wide explanatory architecture that enables the high-level decision making parts of a machine to process the explanatory reasons supporting a decision. The proof-of-concept is designed with reasonableness monitors around each component (including the planner). The system-wide planner monitor examines the planner's proposed plans along with the explanations from the underlying parts to make a more informed, explainable, and robust decision. The architecture also includes a priority hierarchy to enforce individual needs when there are conflicts [3]. Such an architecture has implications to policy and ethics, where the user could designate their needs hierarchy, and understand the factors leading up to a decision.

When autonomous machines get into accidents, as they already have[5], society, lawmakers, and insurance companies will ask why. These autonomous machines should be able to provide evidence in the case of such an accident, similar to how humans provide evidence. For this kind of auditing, I built after-the-fact qualitative reasoning systems that maintain dependencies and construct an explanation of what happened from a log [9]. With the input of a simulated Controller Area Network (CAN) log from a semi-autonomous vehicle, the reasoning system performs edge detection, interval analysis, and event propagation to automatically track the required dependencies, which are used to construct symbolic arguments.

[4] https://www.engadget.com/2018/05/07/uber-crash-reportedly-caused-by-software-that-ignored-objects-in/

[5] https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

Part of those dependencies are causal chains that come from the expectations determined by the model, and some are constraints coming from recorded data. These types of explanations retain honest and clear records of the actions that the agent took, which are suited for accountability and legal reasoning [4, 8].

Research Vision: Articulate Machines

My long-term research vision is that, in the future, complex machines and systems will be **articulate by design**. Dynamic, internal explanations will be part of the design criteria, and system-level explanations will be able to be challenged in an adversarial proceeding. Further, explanations are dynamic: if the explanation is inadequate or inappropriate, the underlying process should be corrected or disabled. With this vision, **all machines will be able to explain themselves**, at various levels of detail. Achieving this vision will require progress in the following three areas: reasoning and representation, narrative and NLP, and human-centered computing. These types of articulate systems are applicable to many domains; I will briefly discuss implications to security and the Internet of Things (IoT).

Hybrid Approaches to Combine Data-Driven Decisions with Symbolic Reasoning and Representations

In critical applications, we impose multiple methods to check and validate our solutions. In the financial realm, we have double-entry bookkeeping, and airplanes have multiple engines and checks for safety-critical components. We should require the machines to also be able to reason about their decisions in multiple ways. But currently, reasoning systems and approaches function in isolation. My work relies on reasoning techniques from commonsense reasoning, case-based reasoning, and hypothetical reasoning. Many of these techniques are used in isolation, but as a first step, I will try to incorporate these approaches together in a hybrid-reasoning system that uses rules, commonsense, and hypotheticals for more robust decision making.

Systems that use higher-level representations are restrictive; they are typically human-curated, static, and specific to the target application or input data. In my reasonableness monitoring system [2], I represented the input descriptions as a composition of conceptual primitives. This representation is difficult to learn automatically. As a first step, I will look at more flexible representations of knowledge and language, and how representations can be learned with limited human-curated information.

Using Explanations as Internal Narrative for Story-Enabled Intelligence

Machines should be able to tell stories like people do. The types of explanations that I develop serve two purposes: they are symbolic to be used internally by the parts of a machine to reconcile their errors, and these symbolic arguments are constructed into a human-readable explanation documenting what happened and why. Processing, understanding, and building these types of explanations is still an open area of research.

Explanations for Society

When autonomous machines share control with a human operator, there will be some explaining to do. If the autonomous operator intervenes, the human will ask why. If the machine operator does not provide a proper explanation, the collaboration will be flawed. They will need to speak a common language, and be able to process, understand, interpret, and intervene based on this language in real-time. I am interested in using explanations, between a human and machine operator for more streamlined and trustworthy decision making. This relies on better design and mechanisms for humans to intervene based on explanations.

Another societal problem is that the legal realm does not support the upcoming transition to autonomous decision making (e.g. AVs). These decision making systems have been showed to cause harm; including racial bias [12] and in some cases, physical harm. In safety-critical and mission-critical decisions, AVs will need to be able to defend their actions, and testify in an adversarial proceeding. As an interdisciplinary direction, I will encode the legal requirements for these machines to be able to explain themselves to abide by legal rules and infrastructure.

Applications for Explainability

Computer security systems are imperfect. Intrusion detection software is good at catching single points of failure and other local vulnerabilities. But most security software fails when there is a faulty connection, an *inexplainable communication* between parts. Using symbolic subsystem explanations can mitigate these intrusions, by consistently and constantly monitoring the reasonableness of subsystem communication and checking the behavior against prior data. This approach is also relevant for IoT systems, where independent entities work together, and enabling common communication will be key to diagnosing errors.

In summary, **autonomous agents are making decisions with consequences**. For stakeholders (e.g. insurance companies, police overseers) and society (e.g. the people harmed) to trust these autonomous thought-partners, the agents need to provide a concise, understandable explanations of their actions. This explanatory ability requires significant technical developments in reasoning, representation, and narrative, while also exploring societal questions in HCI, and policy. In the current interim of shared autonomy (between human and machines), a path of adoption includes the use of monitoring, learning from explanations, and adaptable representations towards system-wide deployment of self-explaining systems. By focusing on limited autonomy, like car that drives itself, agents are no more aware than they are hard-coded to be. Instead, my approach is to create the capability for a complex machine, like a car, to be aware of its internal state as well as its environment; in other words, a car that knows, and can explain that it is driving.

References

[1] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 2017.

[2] Leilani H. Gilpin. Reasonableness monitors. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] Leilani H. Gilpin. Explaining possible futures for robust autonomous decision-making. *To appear in the Proceedings of the AAAI Fall Symposium on Anticipatory Thinking*, 2019.

[4] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[5] Leilani H. Gilpin, Tianye Chen, and Lalana Kagal. Learning from explanations for robust autonomous driving. In *ICML Workshop on AI for Autonomous Driving*, 2019.

[6] Leilani H. Gilpin and Lalana Kagal. An adaptable self-monitoring framework for opaque machines. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1982–1984. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[7] Leilani H. Gilpin, Jamie C. Macbeth, and Evelyn Florentine. Monitoring scene understanders with conceptual primitive decomposition and commonsense knowledge. *Advances in Cognitive Systems*, 6, 2018.

[8] Leilani H. Gilpin, Cecilia Testart, Nathaniel Fruchte, and Julius Adebayo. Explaining explanations to society. *arXiv preprint arXiv:1901.06560*, 2018.

[9] Leilani H. Gilpin and Ben Ze Yuan. Getting up to speed on vehicle intelligence. In *AAAI Spring Symposium Series*, 2017.

[10] Adam Gleave, Michael Dennis, Neel Kant, Cody Wild, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.

[11] Roger C Schank. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631, 1972.

[12] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.