Verification of Autonomous Systems Workshop ICRA 2018 - Session 3 Notes

Don - Trust but verify. Well-studied: human:human, human:pet, etc. Focus on human -> robot (not robot -> human). Barrier certificates. Others - qualification mechanisms.

Robert - Many robots - field roboticist, multi-robot systems. Safety of lives, people, property. Robust ability to do the right thing. Algorithms with analytical guarantees - new systems and questions. Formal methods and temporal logic. Classification of time-series data, anomaly detection, diagnosis, etc. Method: (1) Explain (intent) (2) Express (results) (3) Reason (answer why?). Not classifier in feature space - want simple language but very expressive. Synthesize policy from specification. Intention inference. Maps physical reality into proofs. How to use it. Beyond inference - robots communicate in English via temporal logic specification - guarantees on search tree.

Will - Reinforcement learning (fan of V&V but not really workable in this domain). Very dependent on simulation! Good direction - I have a good guaranteed control system, so use that! Many try to generate control algorithms using supervised learning, but try to not move too far from the math. Note that the current trend towards end-to-end deep learning replacing systems we understand at a theoretical level seems to be taking us AWAY from the ability to V&V our systems, not towards it.

Fabio - V&V applied to deep learning. Benchmarking - robot not hurt people, robot not fall apart. Like a car with very concrete metrics: time until break, speed. Robot - airplane autopilot - easy - series of checkpoints - maintain altitude and speed. Difficult problem - human hand - can it catch objects? What kind? Don't know environment. Metrics - should look at statistical and probabilistic black box measures. Formal methods -> correct software, but don't handle environment - probability that system performs the task set? Trust airplanes even through landing and takeoff. Autonomous cars - highways, instrumented cities. Google's test approach still hasn't resulted in vehicles being released into the wild. Can't have certainty, so provide probability distribution that system will perform. Requires many many many trials. Operational research - develop high fidelity simulations and perform lots of sim runs to validate simulators. What methods and best practices enable trust in the simulator Matrix of probability distributions that characterize system in terms of behavior. Deep learning = black box. Check that it works (target environment, simulation, field). Reproducibility.

Panel:

Don has questions for the panel and/or the audience to drive discussion. (1) necessary, critical for adoption? Role is more than that - drive the theory. Do lot of claims - seldom verify even reproducibility, much less function or properties. Physics (predictable, reproducible) vs. alchemy (not).

Legal and regulatory restrictions driving adoption. FAA not okay with learning systems. Uber gets government pushback, but society wants it and overrides government - demand is driving adoption.

Application of existing verification tools to verify autonomous systems. To the extent we can specify the model and the environment. Other safety critical systems - metrics to determine reliability? Rigorous mechanisms for physical systems and controllers. SIL process (safety integrity level) - predictable, deterministic responses in terms of relationship with humans, but don't have just a few variables in autonomous systems. levels of complexity of the interaction with the environment.

Highest level autonomous systems driving the right thing same as person doing the right thing? Numeric tests for consciousness - know humans well - go by analogy - questionable whether one metric fits other systems.   Limitations on extending autonomy. Making decisions on local environmental information. anti-lock brakes. Simple decision and acting. Complex systems? verify doing right things in many situations? New approaches? Yes! Much increased complexity. How to describe? (language) Agree, but not compare to human performance. Robots not generalists in same way. example: don't know why machine learning did X - response: human occasionally doesn't know either. explainable AI (relationship to verification?) Toolbox of verification tools with incremental testing.

See a time where verification does not involve humans? Yes - meta-cognition work. Descriptive languages. How agile is explainable AI? cultural, societal question determined by level of risk - cost driven by verification process (safety critical systems) - cultural vs. technical problem? When we know what to check, we can automate. Would like to say cultural, but cetrification of humans as drivers - (London taxi drivers) Humans have already proven their ability to drive - robot is new. Culturally right - right now don't care about imperfect things when they're not important (e.g. Alexa vs. Waymo). Care about important things like tractors and cars.

Robots can get worse with learning. How to validate policy space that could be learned? Potential for verification technology to say it won't get worse? Wrong question - focus on how it fails - make sure it's not doing the wrong thing and let the right thing take care of itself. Easier question - place for online verification? NASA redundant backups.