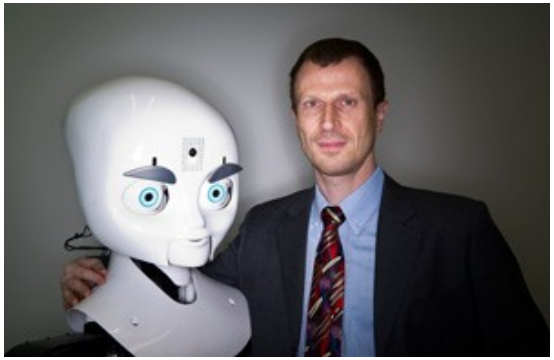
The Need for Explicit Ethical Mechanisms in Architectures for Autonomous Systems [(*)]

Matthias Scheutz
Human Robot Interaction Laboratory
Tufts University
Medford, MA 02155, USA
http://hrilab.tufts.edu/



[(*)] Scheutz, M. (2017). The Case for Explicit Ethical Agents. AI Magazine, 38(4), 57-64.

Summary of the argument

- Dealing only *implicitly* with ethical principles in autonomous agents, e.g., by “considering” ethics somehow subsumed in “reward functions” or by learning human preferences from observations alone is dangerous and needs to be prohibited!
- For algorithms learning from observations alone cannot learn the difference between contingencies and obligations or prohibitions, and thus can also not learn *norm conflicts*
- As a result, such algorithms cannot generate true explanations which require recourse to ethical principles, because they *never learned any principle in the first place*
- For us to be able to **trust** that a machine has understood our ethical principles and acts in accordance with them, we need an agent architecture that can **explicitly** represent them (which we can then verify through inspection) and use them for its decision-making and for justifying its behavior

A wager and false dichotomy

- At AAAI 2015 a panelist posed the following question to the audience: “What would you rather take: an airplane whose controls have been proven correct formally, but which has never flown, or an airplane controlled by a deep neural network that has been flown successfully for 10,000hrs?”
- However, this is comparing apples and oranges
- The correct question is either “would you rather take an airplane with controls proved correct or one with a deep neural network control neither of which have been flown?”
- Or, alternatively, “would you rather take an airplane with controls proved correct or one with a deep neural network control both of which have been flown successfully for 10,000hrs?”
- I would opt for the one with the proven controller in both cases, so do most people I have asked

Is logic is out?

- A frequently heard argument *against* logic-based approaches to ethical behavior conflates provability/guarantees with determinism/stochasticity, i.e., logic is deterministic, the world is stochastic, hence we cannot use logic
- Even for stochastic worlds, however, there can be **provable guarantees** of logical specifications, e.g., take probabilistic model checking which yields “satisfiability of a formula with probability greater than some threshold”
- Another argument is based on the contradictory nature of human norm systems, i.e., our norms are often inconsistent, but inconsistent obligations imply anything is obligated already in the simplest of all deontic logics
- While true, this is not a good reason to abandon logics for reasoning about permissions and obligations, it just means that principles for resolving norm conflicts are needed

Problems with RL and IRL

- **Reinforcement learning** (RL) is a widely used method for learning “optimal action policies”, i.e., ways to act given the state of the agent and the environment using a “reward function” (that encodes the agent’s goals)
- **Inverse reinforcement learning** (IRL) attempts to learn such a reward function from observed or demonstrated behaviors (i.e., sequences of state-action pairs)
- There are two immediately obvious problems with IRL for learning ethical behavior:
  - what if an action *a* was *never* observed in state *s* because it is **prohibited**, will the agent learn the prohibition?
  - what if an action *a* was *always* observed in state s because it is **prescribed**, will the agent learn the prescription?

Problems with RL and IRL

- The answer to both questions is “no”
- Because there is no way for the agent to determine the difference between regularities (that might happen for no particular reason) and normative prohibitions/prescriptions
- Another problem is that observed (human) behavior might be suboptimal or even contain ethical transgressions, in which case the artificial agent will learn them as well
- And while there are various recent methods for learning improved behavior from observations of suboptimal behavior, it is unclear whether the optimization will, in general, respect norms; for one, if a prohibition causes suboptimal behavior (e.g., not entering a one-way street which otherwise would be a shortcut), it likely that the agent will attempt an action that can improve its performance and thus violate the prohibition (and analogous for prescriptions)

Problems with RL

- Another set of problems is connected to generalizability and to prohibitions in light of how RL learns
- For RL systems, the problem is always how to write down the reward function to accomplish compliance with norms (one way would be to give norms instead of reward functions, which is something we have explored)
- Since RL needs to “explore” actions to learn about their utility and effect, it is likely that (without additional constraints) it will try out forbidden actions as part of its exploration and we certainly don’t want our autonomous car to kill a person only to learn that this is not desirable
- Explicitly giving the system all prohibitions and prescriptions for all states ahead of time, however, defeats the purpose of having an RL system learn them in the first place

Problems IRL

- IRL learners will likely not see enough real-world (DoD-type) demonstrations of an ideal autonomous system in all cases that matter (if they can get those observations at all) to learn optimal policies or to make normative generalizations
- While norms provide explicit rules for generalization (across different states), it is unclear **how** IRL systems will generalize from observed behavior (methods for smoothing reward function representations in deep neural networks might run counter to the "reward space" of real-world norm systems)
- And, in general, we want our autonomous system to not perform only at the human level, but ideally be "super-human" norm followers, what philosophers call "supererogatory" (e.g., always **self-sacrifice** in order to save human lives which cannot be imposed on any human), how could an IRL system learn supererogatory behavior?

How to fix the problems

- Instead of learning reward functions or human preferences (which are equally problematic), the agent should learn **ethical principles** (i.e., specified in some formal language)
- In Kasenberg and Scheutz 2017 we proposed a method for learning norms specified in *linear temporal logic* (LTL) from observations that optimized both the formula's complexity as well as the extent to which it explained observed behavior, resulting in the smallest set of LTL formulas that most closely approximated the observed behavior (under the assumption that the observed agent followed those norms)
- The approach utilized a stochastic MDP framework for intrinsically dealing with **norm conflicts** based on an algorithm that **provably** suspends the smallest number of conflicting norms for the shortest time to allow the agent to obey the remaining norms based on weights or priority orderings (see Kasenberg and Scheutz 2018)

How to fix the problems

- Equipped with the representational repertoire for specifying temporally extended norms, the agent can introspect on its norms and can use them in explanations and justifications of its behavior (an RL/IRL agent can at best report the “Q value” of a state-action pair, but cannot make recourse to principles it has never learned and does not represent)
- In Kasenberg et al. 2019 and 2020 we show how our agent can genuinely answer “why” questions about past events as well as hypothetical and counterfactual events (and not just concoct post-hoc “interpretations” of deep neural nets)
- Counterfactual answers are particularly useful for us to understand decisions and tradeoffs with norm conflicts: “why did you do action A and not action B in context C?”
- In this case, the system can show how, had it done B, the outcome would have been *worse in terms of violating norms*

Conclusions

- Implicitly learning how to behave ethically (through RL or IRL methods, which adherents of “value-alignment” aim for) is riddled with problems and will not accomplish what we need for autonomous systems, especially in DoD contexts
- Having explicit logic-based representations enables explicit instructions of norms (which is inevitable because most such principles cannot be learned from observation alone)
- It also enables the system to use norms directly for making decisions and for generating explanations (these are genuine explanations because they are produced by the same algorithm the system used to generate its behavior)
- Finally, in cases of norm conflicts we require justifications to make explicit recourse to principles for us to be able to understand the tradeoffs (trivially, systems that never learn such principles cannot use them to generate justifications)

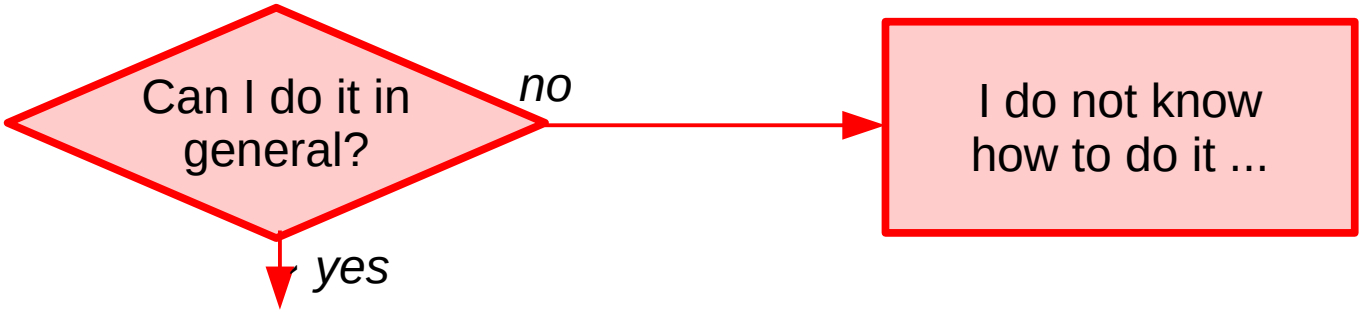
Acknowledgments and links

- Students/staff: Daniel Kasenberg, Ravenna Thielstrom, Thomas Arnold, and several other students in the HRILab
- Funding provided by NSF, ONR, and DARPA

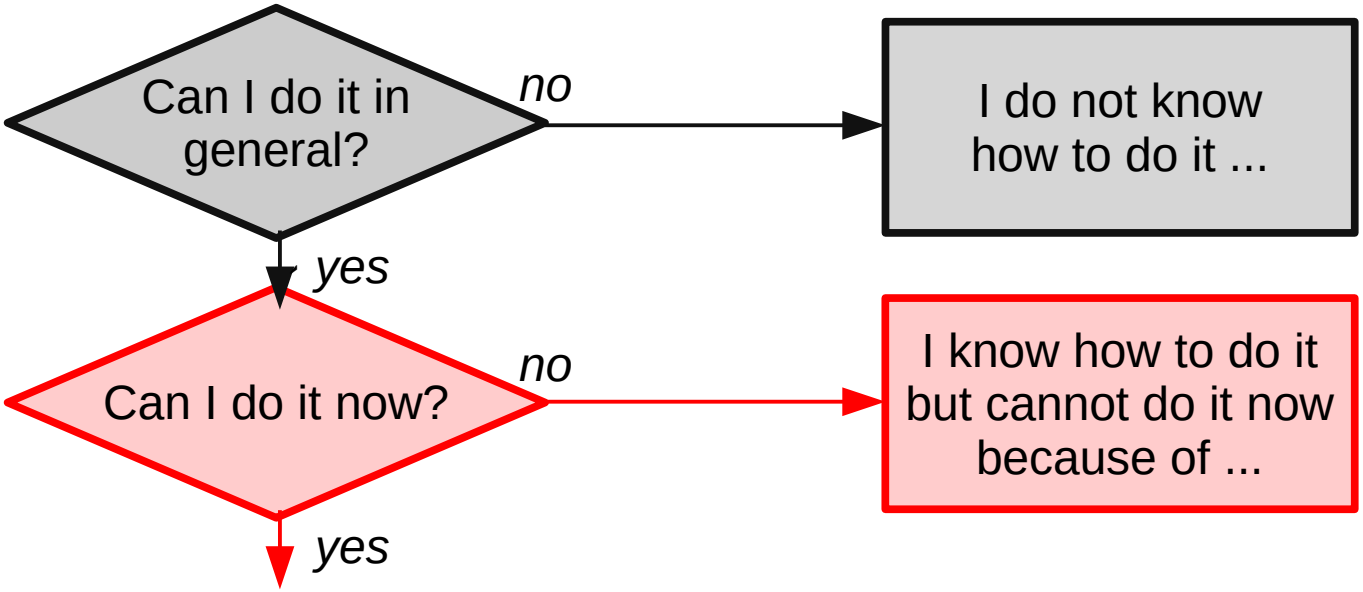
- References:

  - D. Kasenberg, R. Thielstrom, and M. Scheutz (2020). “Generating explanations for temporal logic planner decisions”, *Proceedings of ICAPS* http://hrilab.tufts.edu/publications/kasenberg2020icaps.pdf
  - D. Kasenberg, A. Roque, R. Thielstrom, M. Chita-Tegmark, and M. Scheutz (2019). “Generating justifications for norm-related agent decisions”, *Proceedings of INLG* *http://hrilab.tufts.edu/publications/kasenberg2019inlg.pdf*
  - D. Kasenberg and M. Scheutz (2018). “Norm conflict resolution in stochastic domains”, Proceedings of AAAI. https://hrilab.tufts.edu/publications/kasenbergscheutz18aaai.pdf
  - D. Kasenberg and M. Scheutz (2017). “Interpretable apprenticeship learning with temporal logic specifications”, *Proceedings of CDC* https://hrilab.tufts.edu/publications/kasenbergscheutz17cdc.pdf

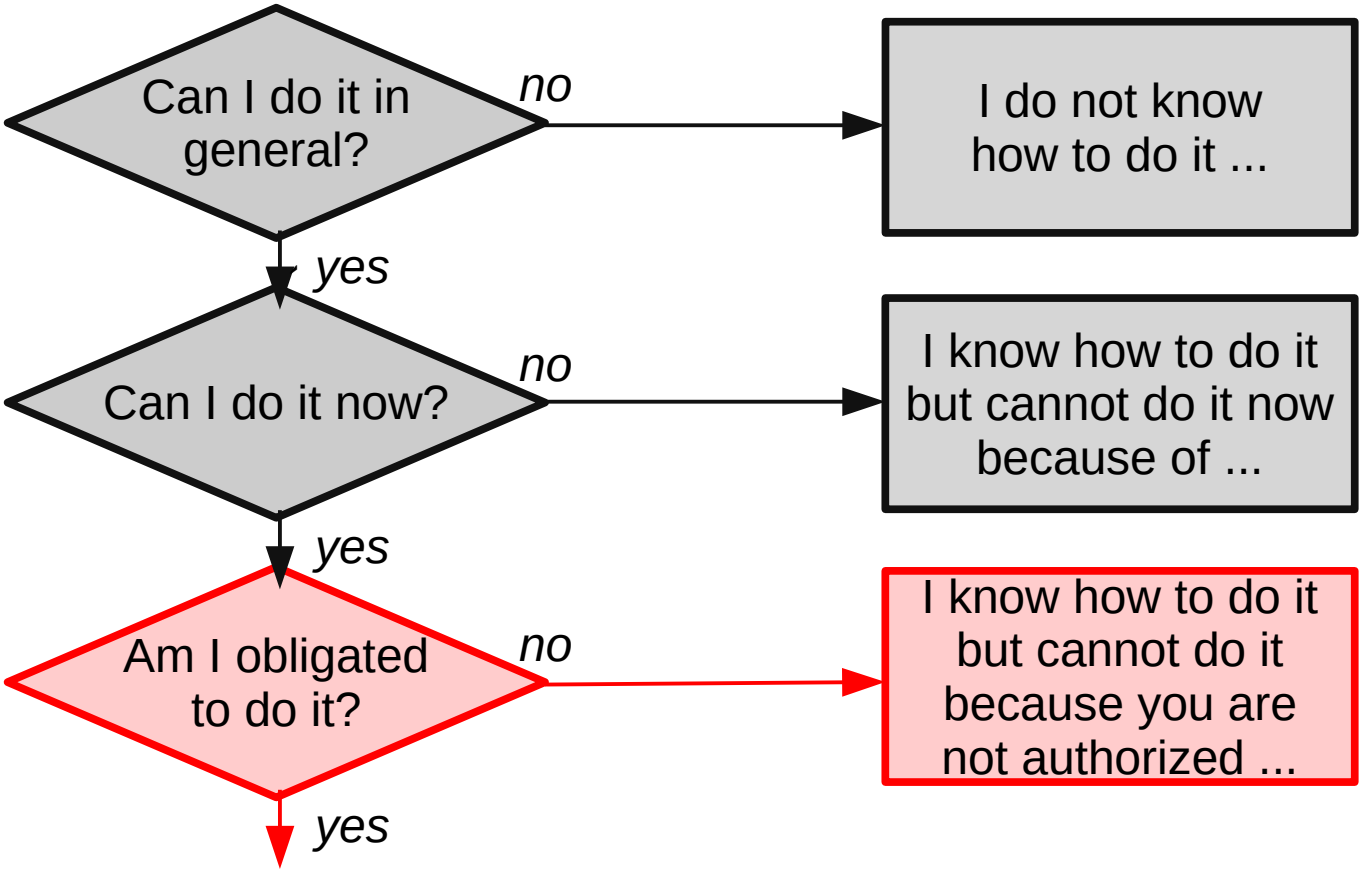
Reasoning about capability



Reasoning about availability



Reasoning about obligation



Reasoning about ethical violations