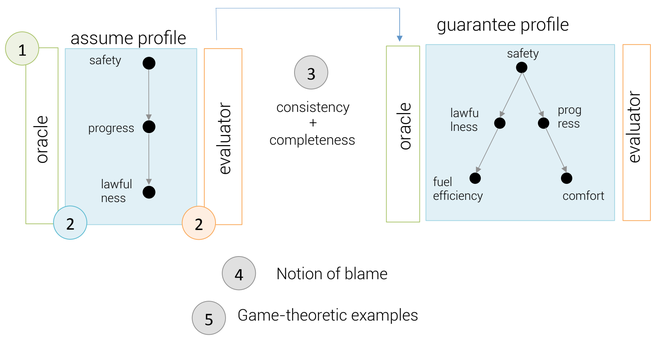
Towards Assume-Guarantee Profiles for Autonomous Vehicles

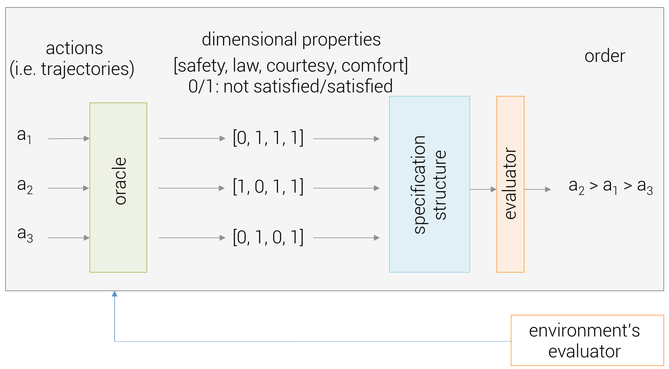
PI: Richard M. Murray
Graduate students: Tung Minh-Phan* and Karena X. Cai*
NSF Award Number 1545126

* Equal contribution

high-level overview of current work



single-agent profile

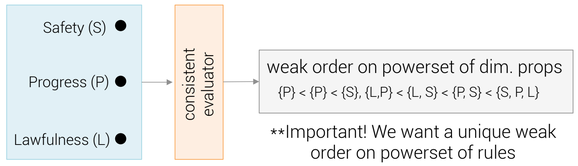


Definition: dimensional properties
A dimensional property is a desirable attribute that must be satisfied or not satisfied. (ex. Safety, lawfulness, courtesy)

Definition: oracle
Abstraction of the self-driving car's perception system.

consistent evaluator

A class of functions that can endow some partially-ordered sets (poset) with a unique weak order on their powersets



**Important! We want a unique weak order on powerset of rules

Definition: weak order

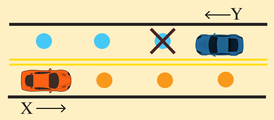
A ranking of a set, some of whose members may be tied with each other (less strict than a total order.

Question:

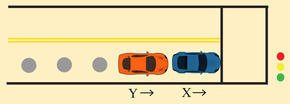
Why ***weak*** instead of ***total*** order?

Motivation & Goals

To construct a high-level controller that will guarantee correct behavior in all situations that arise.



1. Create a set of axioms for a self-driving cars for which the specifications are both complete and consistent.
2. Make the car decision-making process transparent.
3. Make reasonable assumptions about other agents in the environment to guarantee correct behavior.



Question: when can a poset be consistently evaluated?

Theorem 1: A finite poset P of dimensional properties has a consistent evaluator if and only if it can be partitioned into a set A of N maximal antichains such that:
1) the maximal antichains A can be assigned ranks in such a way that the partial order is respected
2) For each dimensional property, there exists a maximal chain containing it of length N

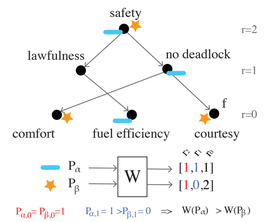
Consistent evaluator function (i.e. properties 1-5 hold)

⇕

Decomposition of poset into maximal antichains with some additional properties

the W function (⇐)

A consistent evaluator on a consistently-evaluable poset.



Theorem 2: Such a partition in Theorem 1 is unique.

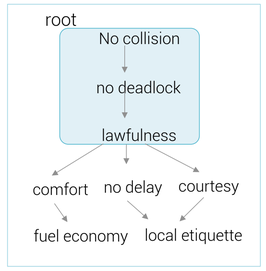
Consistency and completeness

Definition: consistency
There is a unique weak-order on the powerset of a specification structure regardless of the consistent evaluator being used.

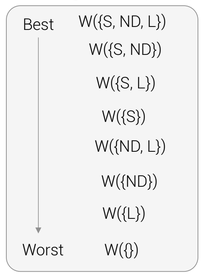
Definition: completeness
More dimensional properties (nodes) = more complete

Simple Example

refinement



ordering of root (via W)



interpretation

1) Best to satisfy all dimensional properties.
2) W({S}) > W({L}) If being safe requires breaking the law, you should do so.
3) W({S, ND}) > W({S, L}) If you can break out of a deadlock situation but you have to break the law, you should do so.

assume-guarantee profiles

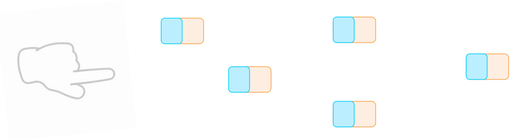
Assumption profiles:
$\mathcal{A}$ a set of behavioral preferences or characteristics that the agent assumes the agent to have

Guarantee profile:
$\mathcal{G}$ a set of behavior preferences or characteristics that it is obligated to behave according to as long as its environment makes decisions in accordance with $\mathcal{A}$

notion of blame

Definition: compatible set
Given $C_j = (A_j, G_j)$, where j is the index of an agent and A_j are the assumptions that agent j is making about its environment while G_j is its guarantees, we say that a group of agents J are compatible if:

$$\forall j \in \mathcal{J}.\forall i \in \mathcal{J} - \{j\}.G_j \subseteq A_i$$

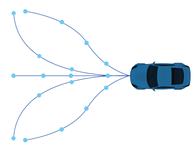


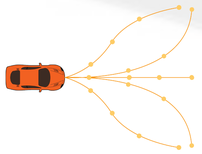
Definition: blame
Assuming all agents are compatible, a blameworthy action/strategy is one in which an agent violates its guarantees, thereby causing another agent's assumptions not to be satisfied and thus resulting in an unwanted situation in which blame must be assigned.

Preliminary Game-Theoretic formulation

N Agents: each agent has a set of state-dependent motion primitives

$$(A^i(s))_{s\in S})_{i\in\{1,\dots,N\}}$$
$$a = (a_1, a_2, \dots, a_N)$$





Rules: cost function ordered according to assume-guarantee profile

$$r^i(SS, s, a)$$

Sequential game: need to reason about joint actions according to some time-horizon

$$v^i_\pi(s) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r^i_\pi(s_t)]$$

Can we solve for the Nash-Equilibria?

If rules in assume-guarantee profile always prioritizes safety above all a, can we guarantee percentage of collisions below certain threshold?