**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 the maximal antichains A can be assigned ranks in such a way that the partial order is respected.

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**

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

**Consistency and completeness**

**Definition:** completeness

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

**Definition:** dimensional properties

More dimensional properties (nodes) = more complete

**Definition:** oracle

Abstraction of the self-driving car’s perception system.

**Definition:** weak order

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

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

1. **assume-guarantee profiles**

   Assumption profiles:
   A set of behavioral preferences or characteristics that the agent assumes the agent to have

   Guarantee profile:
   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 it.

2. **Preliminary Game-Theoretic formulation**

   **Notion of blame**

   Definition: compatible set

   Given C = (A_j, B_j) where A is the index of an agent and A, B are the assumptions that agent j is making about its environment while B_j is its guarantees, we say that a group of agents J are compatible if

   **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.