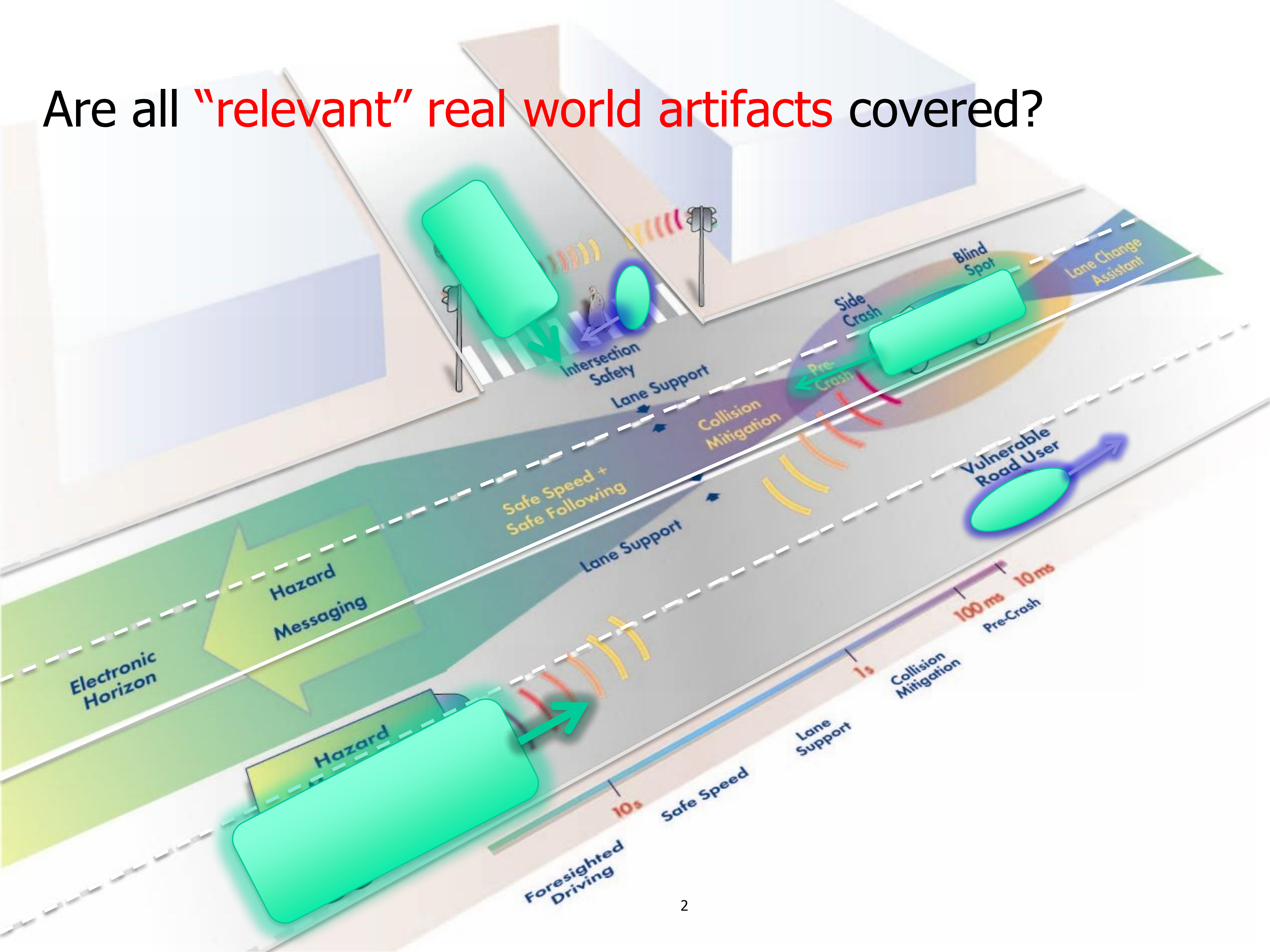
Does It Pay to Extend the Perimeter of a World Model?

Werner Damm[1] & Bernd Finkbeiner[2]

[1]Carl von Ossietzky Universität Oldenburg [2]Universität des Saarlandes

Transregional Collaborative Research Center AVACS

Are all "relevant" real world artifacts covered?

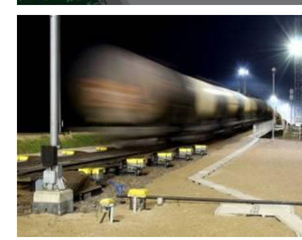


Embedded systems

Embedded system are similar to humans: they

- “observe”
- “analyse”
- “decide”
- “act”



Key to their operation is the capability to “reconstruct” an internal representation of the real world – a world model

Questions

- "observe"
  - are all "relevant" real-world artifacts part of my world model?
  - can the system observe all "relevant" real-world artifacts
  - can we characterize (formally) the notion of "relevance"
  - is there a notion of optimal world models?
- "analyse"

  the possible moves of the adversary (the environment): can they block my objectives?
- "decide"

  give strategy which, based on previous observations, decides how to
- "act"

The discrepancy between the real world and what the aircraft perceives as real decide over life and death

UNIVERSITÄT DES SAARLANDES
ALBERT-LUDWIGS UNIVERSITÄT FREIBURG

14.09.1993 -
Aircraft thought it was still airborne, because only two tons weight lasted on the wheels due to a strong side wind and the landing maneuver. The computer did not allow braking.
The plane ran over the runway into a rampart.

THE SYSTEM ENGINEERING CHALLENGE

Given
a (physical) system S under development

what real-world aspects
could potentially impact S
in a way that endanger its proper functioning?

Industrial Practice: learning processes

- Company XY
  - all flight incidents are analyzed
  - to identify the process step in which the potential for an incident should have been detected
  - existing models are extended to allow the prediction of such potential incidents
  - measures protecting against such hazards are integrated into the design (and aircrafts)
  - safety processes are used to demonstrate resilience against root cause for such hazards

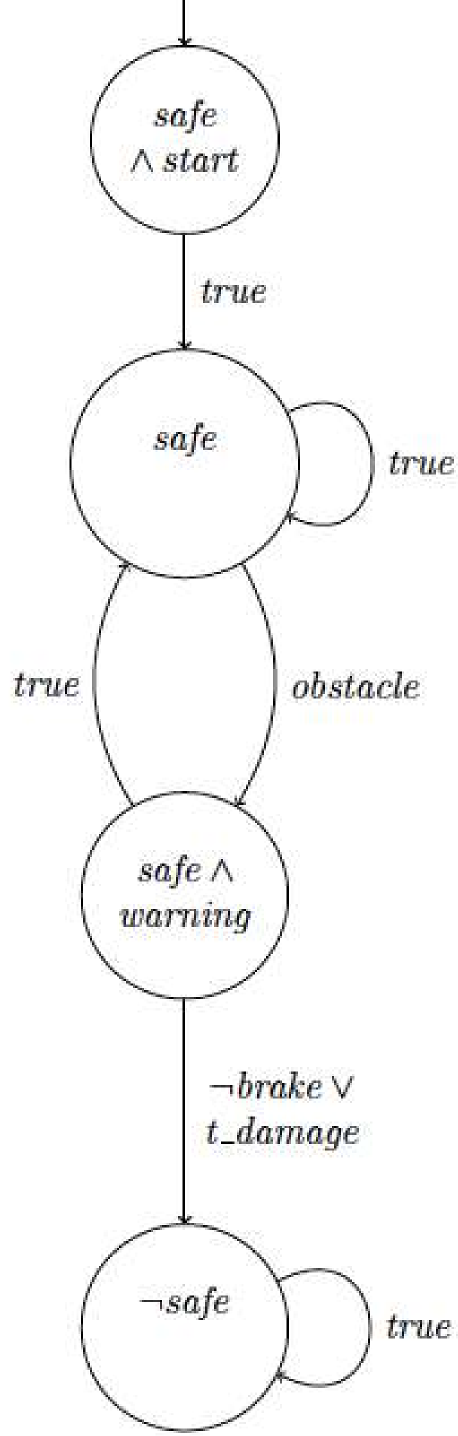
From yes/no to: could we do better?

universität OLDENBURG
ALBERT-LUDWIGS UNIVERSITÄT FREIBURG
UNIVERSITÄT DES SAARLANDES

- No world model will ever be complete
- Hence no formal verification of a cyber physical system can "guarantee" safety (e.g. no crash)

We *"measure" the benefit of extending a world model W*
to include a new real world artifact a
by comparing the strategic capabilities of W and W∪{a}:

Does the richer world model allow to define strategies,
which, in comparable environment moves,
*allow **more often** to achieve the systems objective*?



A simple world model

... for an ADAS to maintain safe distance to objects ahead on same lane (cars, cargo, ...), two lane hwys, secondary objective avoid braking

disturbances

appearance of an obstacle

tire-burst

controllable actions

brake

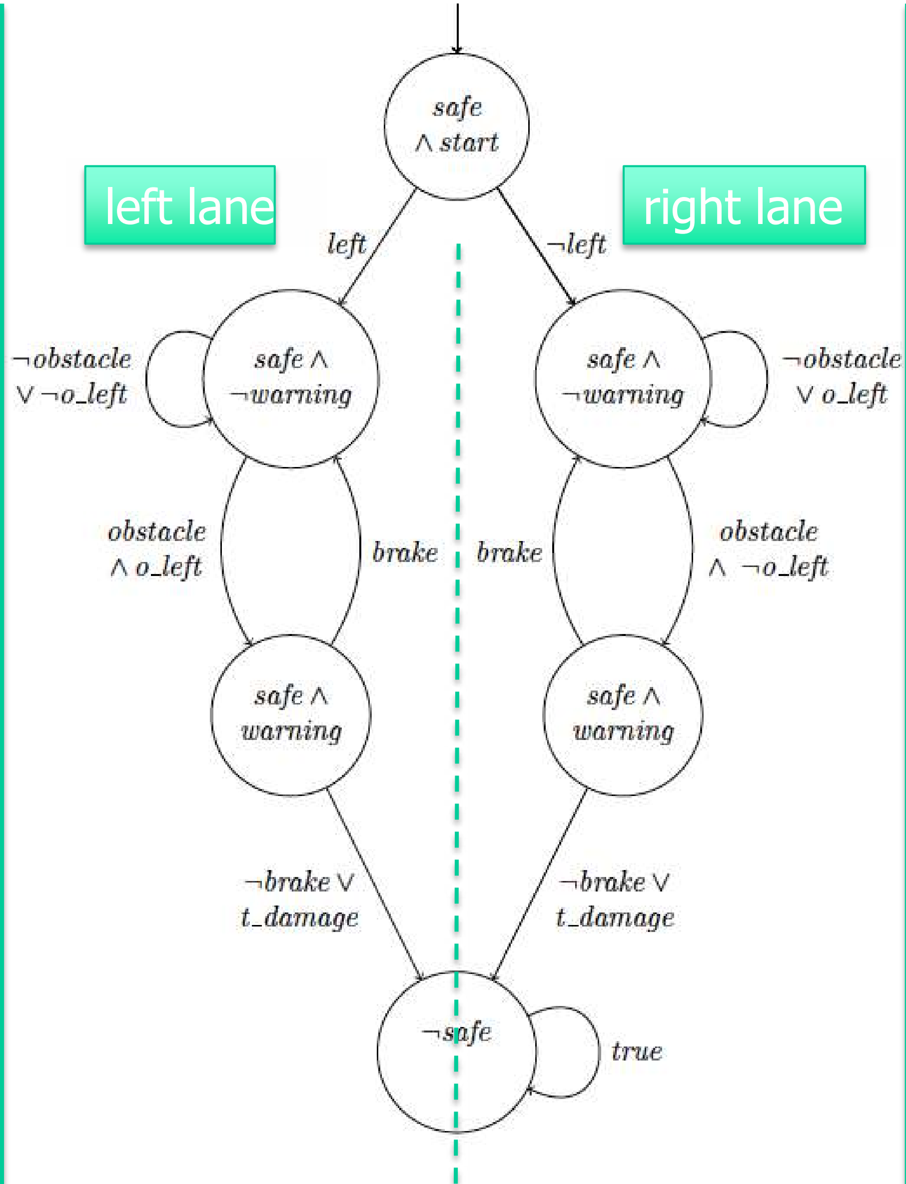
states

safe: the distance to the object ahead of the ego car is greater than some constant

warning: an obstacle has been detected ahead of the ego car

The world models explains how it changes state depending on disturbances and controllable actions

A richer world model



disturbances

- appearance of an obstacle
- o_left: on left lane
- tire-burst

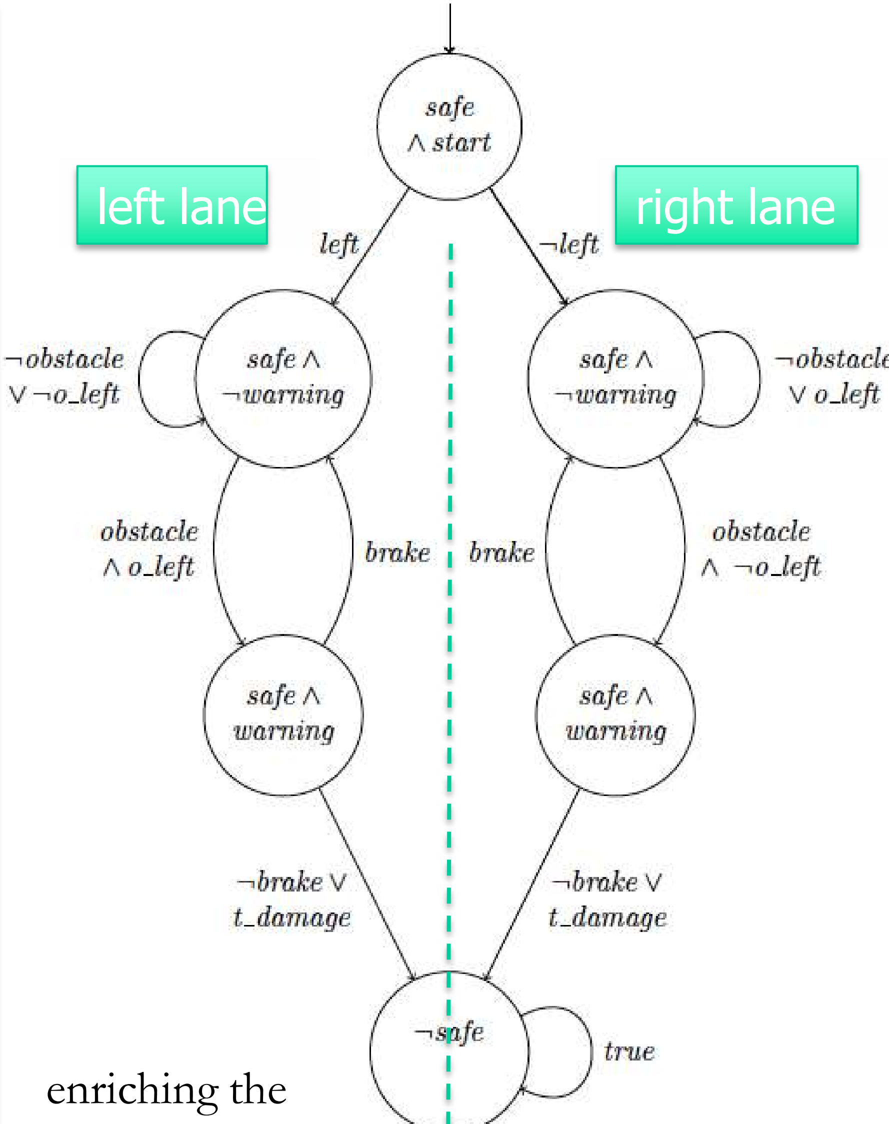
controllable actions

- brake
- left: take left lane

states

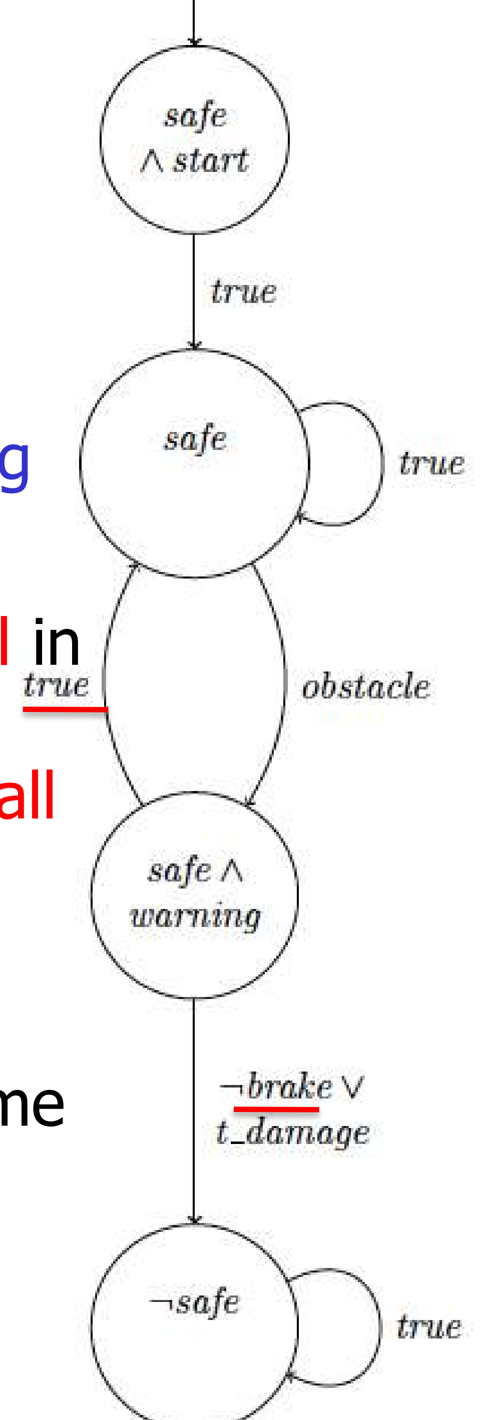
- safe: the distance to the object ahead of the ego car is greater than some constant
- warning: an obstacle has been detected ahead of the ego car

Beyond YES/NO



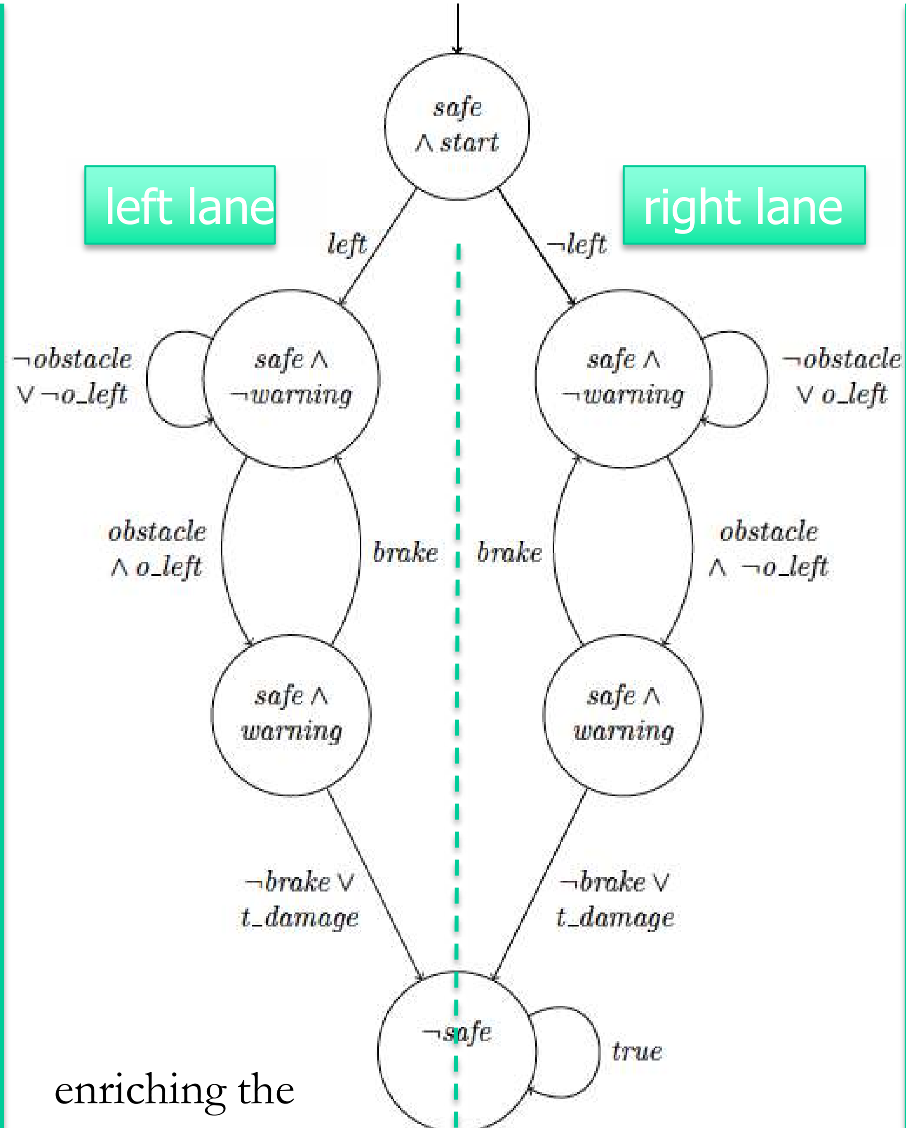
1. Never brake
2. Brake iff warning
3. Brake always

- all strategies fail in both models to always achieve all objectives: tire damage can always cause system to become unsafe



enriching the world model eliminates non-determinism

Comparing strategies: remorse-free dominance



enriching the world model eliminates non-determinism

s_1 Never brake

s_2 Brake iff warning

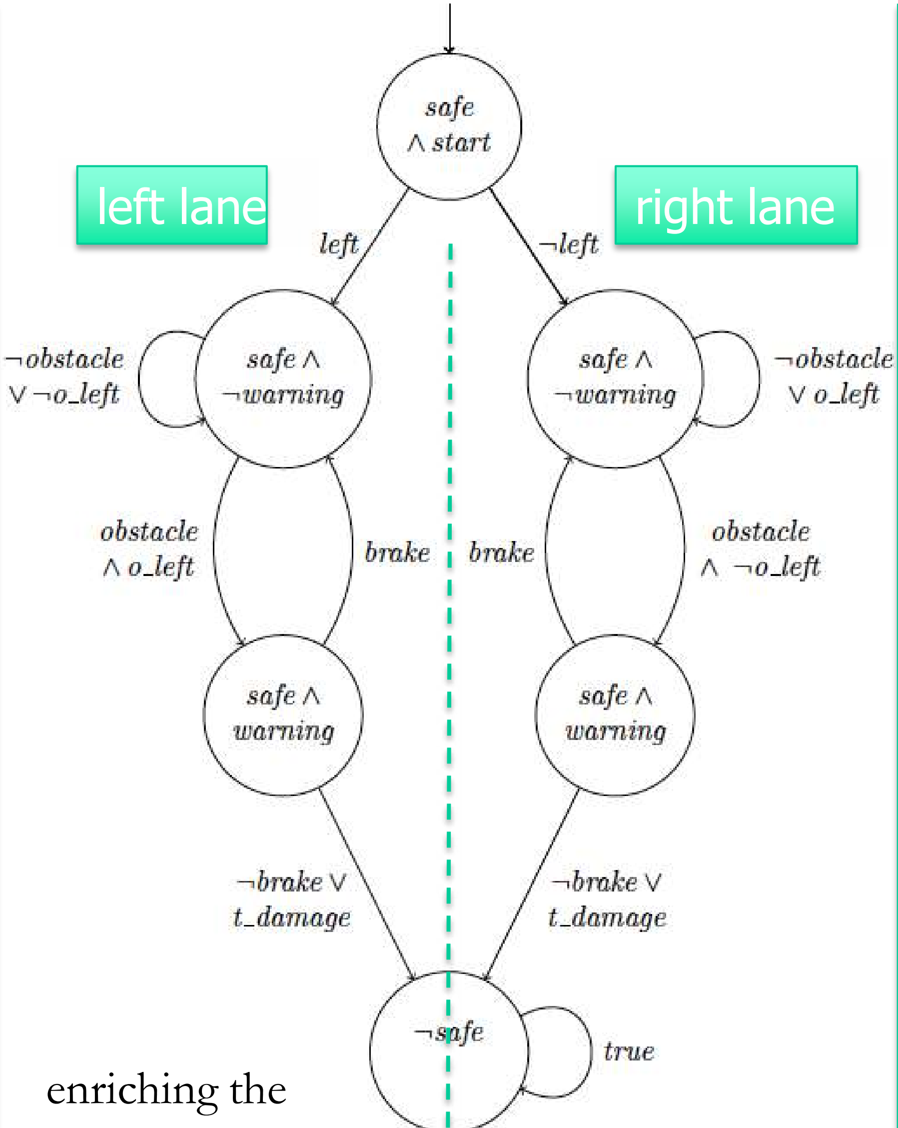
s_3 Brake always

- compare strategies wrt remorse: *could I "have done better"* = achieved higher priority objectives

in "comparable situations" = same sequence of disturbances

- s_2 dominates s_3:
  - whenever s_3 achieves up to prio_x in some sequence of disturbances, so will s_2
  - but s_2 avoids (unnecessary) braking in safe state with no warning

Comparing strategies: remorse-free dominance



enriching the world model eliminates non-determinism

s_1 Never brake

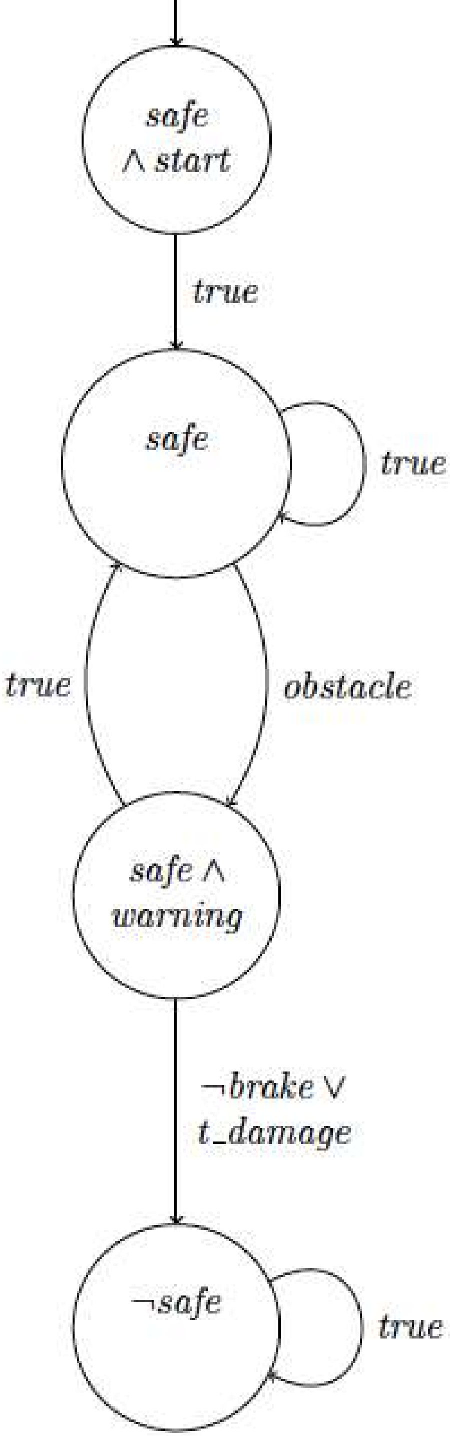
s_2 Brake iff warning

s_3 Brake always

- compare strategies wrt remorse: *could I "have done better"* = achieved higher priority objectives

in "comparable situations" = same sequence of disturbances

- s_2 dominates s_1:
  - whenever s_1 achieves up to prio_x in some sequence of disturbances, so will s_2
  - but s_1 can cause crash in sequences of disturbances where s_2 will remain safe



s_1 Never brake

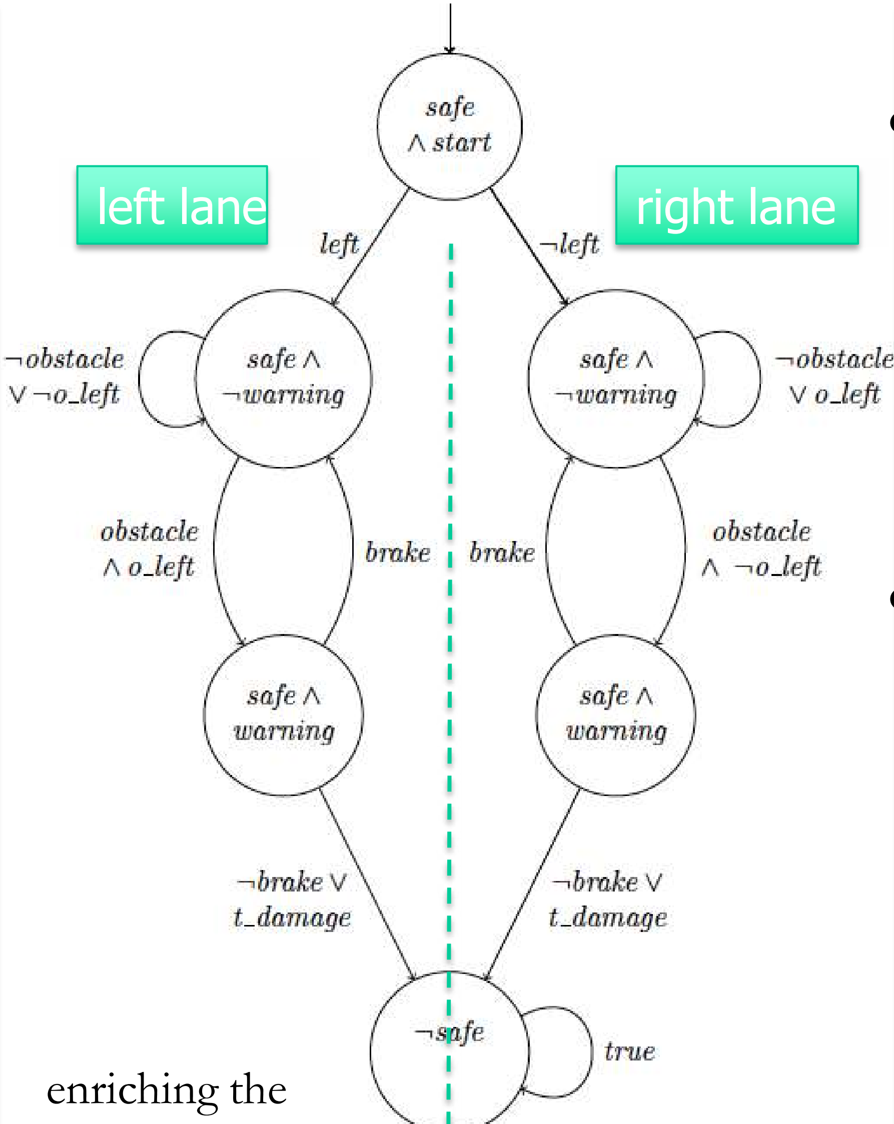
s_2 Brake iff warning

s_3 Brake always

- s_3 is not dominant, because it brakes even in the middle safe state, where there is no danger for safety (hence braking is causing remorse because both s_2 and s_1 avoid this)
- s_1 does not dominate s_2 , because it does not avoid crashes in sequences of disturbances, where this is avoided by s_2
- s_2 does not dominate s_1, because for some sequence of disturbances braking is not necessary to avoid crash (if obstacle is on other lane)

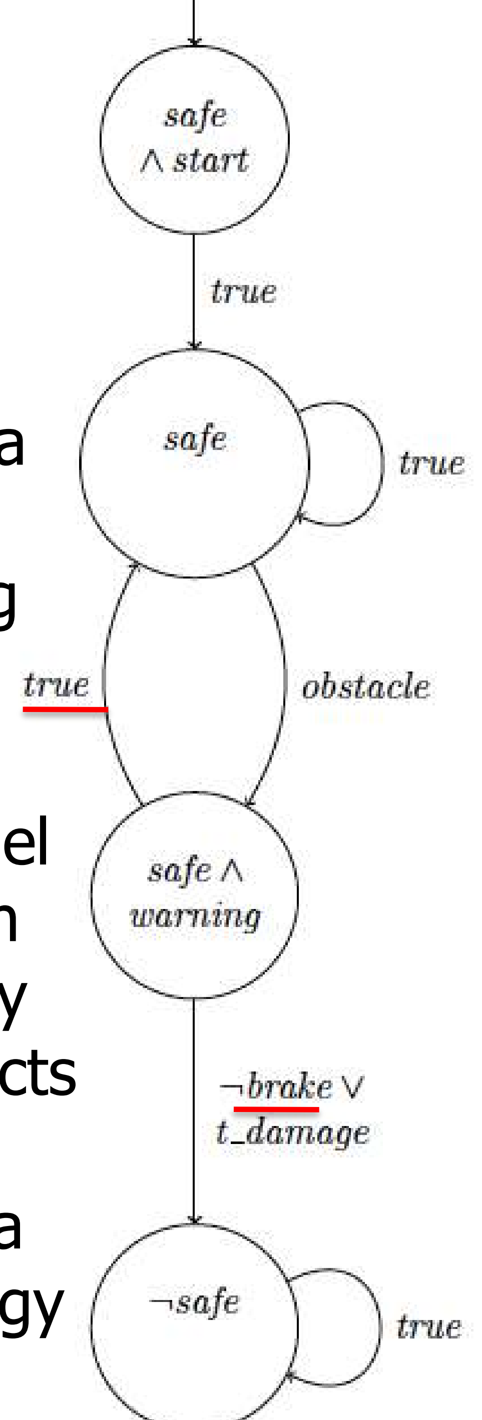
The simple world model does not permit a dominant strategy

It paid to enrich the world model



enriching the world model eliminates non-determinism

- In the refined model, there is a “best in class” strategy: picking this will never cause remorse
- The simple model does not contain sufficiently many real world artifacts so as to allow construction of a dominant strategy



Optimal world models

- Intuitively, given a fixed set of prioritized objectives, only a subset of all real world artifacts are required to define the "best possible strategy" for these objectives

  We formalize this intuition as follows:

A world model W is optimal
if it allows do define a ("best") strategy
which not only dominates all other strategies in W,
but also those definable in all refinements of W

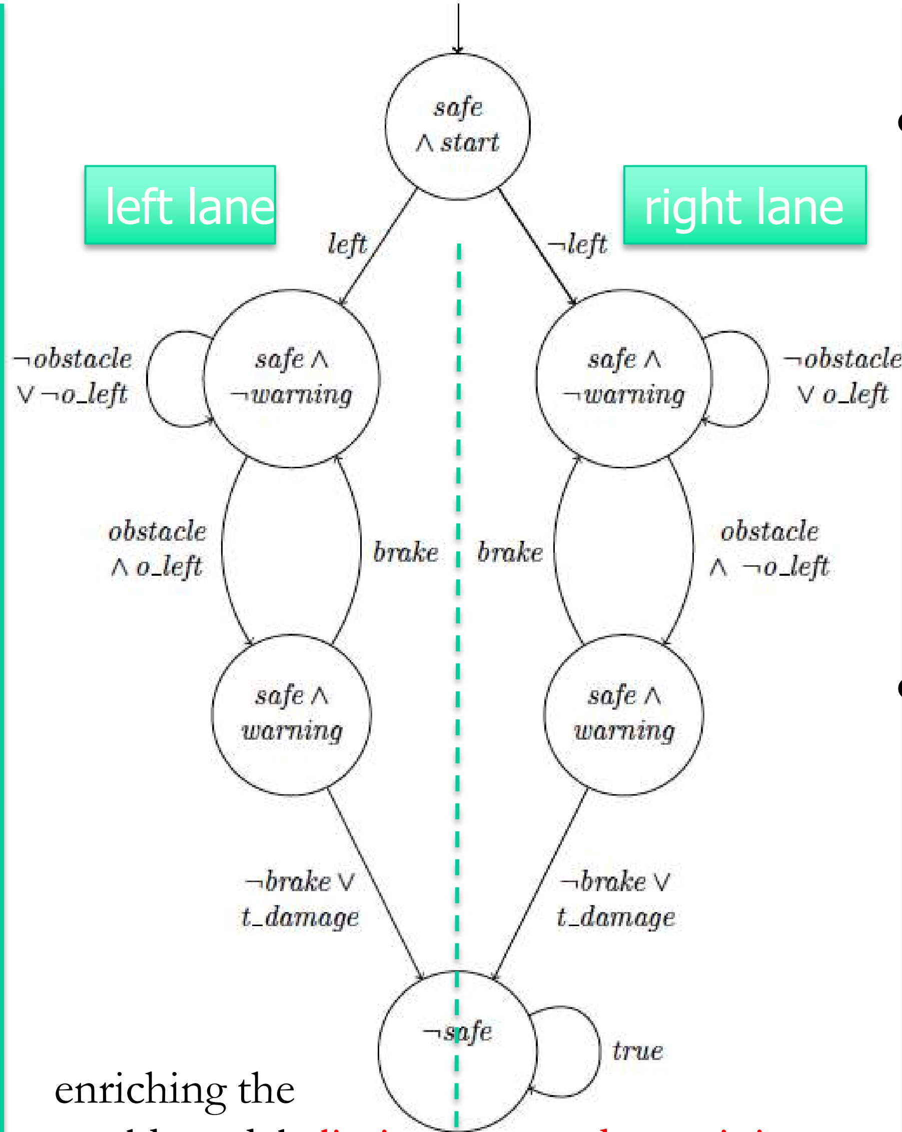
Optimal world models

A world model W is optimal
if it allows do define a ("best") strategy
which not only dominates all other strategies in W,
but also those definable in all refinements of W

Theorem
Let W be a world model, ϕ an objective specification

(1) We can automatically check whether W is optimal for ϕ
(2) If true, we can automatically synthesize a „best" strategy

One more dimension: sensors



enriching the world model eliminates non-determinism

- Its in the model, but does the system "see it"?
  - We designate a subset of the variables of the world model as "observable" (corresponding to sensors) and index strategy classes by observables allowed for decision making
- Does it pay to include additional sensors?
  - No, if it does not help in avoiding remorse,
  - i.e. if there is a strategy with restricted observability dominating all strategies with richer observability

Does adding sensors pay?

Theorem

Let W be a world model, ϕ an objective specification and S_{I1} and S_{I2} two strategy classes over W with observables $I1 \subseteq I2$.

(1) We can automatically check whether a given strategy in S_{I1} dominates S_{I2}

(2) We can automatically check whether S_{I1} contains a strategy that remorsefreely dominates S_{I2} (and synthesize it)

Optimal world models wrt given sensors

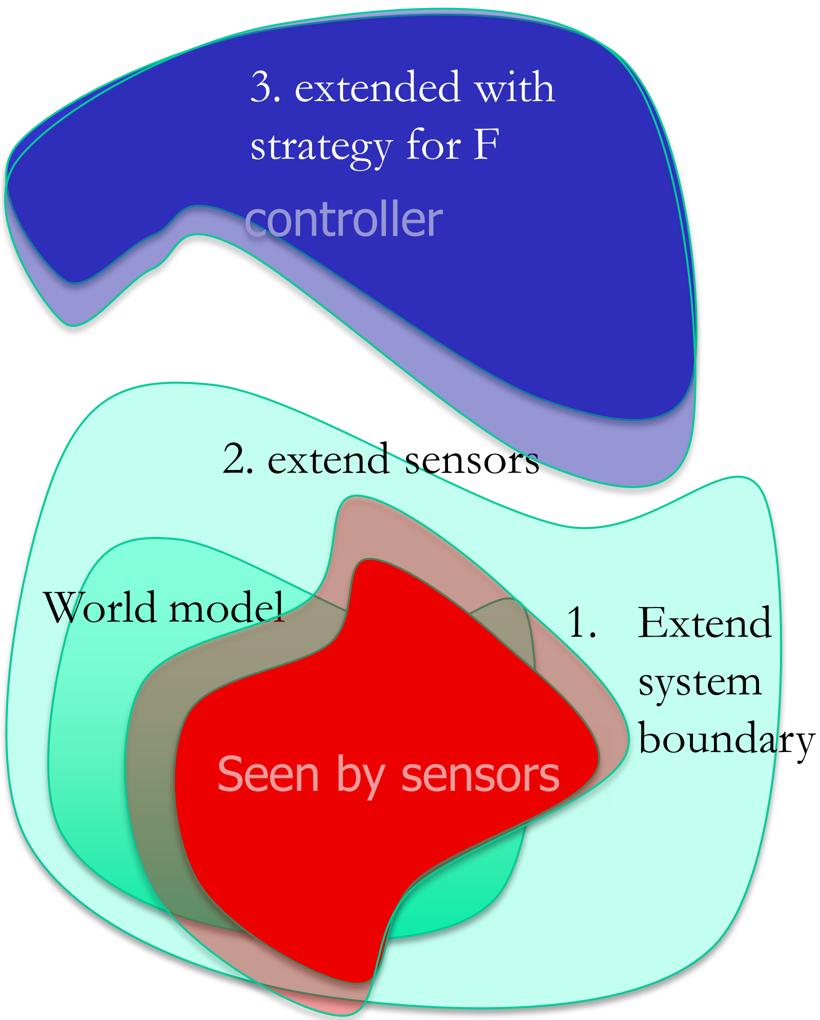
Theorem

Let W be a world model, ϕ an objective specification, and S_I the class of strategies over W with observables I

(1) We can automatically check whether W is optimal for ϕ and S_I

(2) If true, we can automatically synthesize a „best" strategy in S_I

A possible design flow (automotive)



Assume that new Feature F is to be included in new car model. Previous car model comes with world model W and sensors I.

1. Specify list of prioritized objectives.
2. Check if W is still optimal, otherwise extend system boundary until optimal model W_F is found, assuming full information, by reducing non-det.
3. Check if restricting observables to I destroys optimality.
4. If yes, add sensors until optimal W_F with new set of sensors I_F is optimal

Conclusion

- We propose a new quality measure for world-models:

  Do they allow construction of strategies, which will never cause remorse wrt to any other strategy in this and all refined models?
- Rather than analyzing correctness (which is un-achievable) we are optimizing for "best effort" (least remorse)
- There is a tradeoff between
  - The degree to which prioritized objectives are satisfied
  - The price to avoid such remorse situations
- In its most critical form, this entails the decision about the price spend to avoid sacrificing critical safety requirements such as guaranteeing collision avoidance (see ALARP principle)
- Our research is a first step towards a systematic assessment of such trade-off decisions in early phases of system design