PURSS: Towards Perceptual Uncertainty Aware Responsibility Sensitive Safety with ML

Rick Salay,[1] Krzysztof Czarnecki, [1] Ignacio Alvarez, [2]
Maria Soledad Elli,[2] Sean Sedwards,[1] Jack Weast [2]
[1]Dept. Electrical and Computer Engineering, University of Waterloo,
Waterloo, Canada {rsalay, k2czarne, sean.sedwards}@uwaterloo.ca
[2]Intel Corporation, Automated Driving Group { ignacio.j.alvarez, maria.elli, jack.weast}@intel.com

Abstract

Automated driving is an ML-intensive problem and its safety depends on the integrity of perception as well as planning and control. Responsibility Sensitive Safety (RSS) is a recent approach to promote safe planning and control that relies on perfect perception; however, perceptual uncertainty is always present, and this causes the possibility of misperceptions that can lead an autonomous vehicle to allow unsafe actions. In this position paper, we sketch a novel proposal for a formal model of perception coupled with RSS to help mitigate the impact of misperception by using information about perceptual uncertainty. The approach expresses uncertainty as imprecise perceptions that are consumed by RSS and cause it to limit actions to those that support safe behaviour given the perceptual uncertainty. We illustrate our approach using examples and discuss its implications and limitations.

1 Introduction

Perception in Automated Driving (AD) relies heavily on Machine Learning (ML) and includes tasks such as classification, object detection, semantic segmentation and object tracking. A significant source of safety risk in AD is due to *perceptual uncertainty* (Czarnecki and Salay 2018) that leads to misperceptions resulting in unsafe actions. However, the ability to accurately estimate perceptual uncertainty can potentially reduce the negative impact if it is effectively integrated into the AD planning and control policy.

Responsibility Sensitive Safety (RSS) (Shalev-Shwartz, Shammah, and Shashua 2017) is a specification for AD planning and control that develops a safety model based on formalising common sense principles for safe driving and by explicitly assigning responsibility for safety to the appropriate road user. The result is a set of detailed driving rules to deal with different driving scenarios. For example, in the "safe longitudinal distance, same direction" scenario, the rear car c_r (ego vehicle) is travelling on a single lane road with another car c_f in front. RSS mandates that c_r is responsible for keeping a safe distance d_{min} from c_f where d_{min} is expressed as a formula in terms of the velocities of c_r and c_f and *general* parameters specifying vehicle response time (ρ), max acceleration ($a_{max,accel}$), and max/min braking ($a_{max,brake}, a_{min,brake}$). If, for whatever reason, d_{min} is breached, this is defined as a *dangerous situation* and c_r is then restricted to actions specified as a *proper response* to the dangerous situation. A proper response is an action that has lowest risk for the given setting of RSS parameters. In this scenario, the proper response, given that t_b is the time the situation became dangerous, is that during the interval $[t_b, t_b + \rho]$, c_r may accelerate any amount not exceeding $a_{max,accel}$ and then must apply at least $a_{min,brake}$ until the situation is no longer dangerous or it has come to a stop. RSS provides rules (in this scenario) for what is consider a safe driving behaviour.

Unsafe behaviour in an AD system may occur either because of (1) misperceptions due to perceptual uncertainty in the perception subsystem that causes the planning and control subsystem to take an unsafe action, or (2) flaws in the planning and control subsystem that cause it to take unsafe actions even when the perception subsystem is behaving correctly. RSS addresses case (2) by providing a specification for safe planning and control, but it relies on perfect perception and thus is susceptible to case (1). The RSS paper (Shalev-Shwartz, Shammah, and Shashua 2017) considers the issue of perceptual uncertainty at a high-level by assuming that the perceptual system is Probably Approximately Correct (PAC), where the behavioural error due to perceptual uncertainty is probabilistically bounded. While this has conceptual appeal, it is not integrated with the RSS rules and remains at a level of abstraction that makes it difficult to see how it can be used in practice.

In this position paper, we sketch a proposal for addressing case (1) using a formal, but pragmatic, approach to analysing and mitigating the safety impacts of misperception in RSS due to perceptual uncertainty. We start by defining a generic framework for representing AD perceptual states and discuss how particular AD implementations must refine RSS perceptual states. Then we use this framework to formally characterise the conditions under which misperceptions cause violations of RSS rules, and therefore, safety risk. Finally, we show how to pragmatically incorporate operational perceptual uncertainty into RSS and how this can be used to mitigate safety risk due to misperception. We do this by first showing how to represent perceptual uncertainty by *lifting* perceptual states to imprecise states called *under*-perceptions and then showing how these can be used with correspondingly lifted RSS rules to handle imprecise states in a sound and safe way.

2 Perceptual Framework

An AD system consists of a perception subsystem producing representations of the world, which feeds a planning and control subsystem that we assume conforms to RSS and decides how to act on these perceptions. The representation of the world state produced by the perceptual subsystem is called a *world model* (Czarnecki 2018a). A world model consists of a set of typed entities (e.g., road users, static objects, situations, etc.) whose states evolve in time.

A *world model schema* defines the structure of world models. We assume this is based on an ontology for driving (e.g., (Czarnecki 2018b)) and has an object-oriented program style consisting of a hierarchy of entity *class definitions*. In particular, classes define *attributes* (e.g., `pos`, `vel`, etc.), computations of *derived attributes* (e.g., computation of d_{min}), *parameters* used by RSS (e.g., $\mathtt{a}_{\mathtt{max,brake}}$) and *behavioural specifications* defining feasible evolutions of a class instance in time (e.g., state machines, differential equations, temporal logic properties, etc.). We assume an entity can change class over time. For example, a cyclist "turns into" a pedestrian when they dismount and start walking with their bike. The set $\mathcal{S}_{\Sigma}$ of world models defined by a world model schema Σ are called the concrete or *precise* world models.

In this model of perception, we assume the Markov property relative to the world models — i.e., decisions can be made based on the current world model without reference to previous world models. However, historical information about an entity (including its relationships to other entities) can be stored as part of its current state if this is needed for deciding actions. For example, when cars arrive at an intersection with a 4-way stop, the right-of-way is given to the car who arrived first (or to the one to the right if both arrived at the same time). Determining whether it arrived first requires remembering its previous position and velocity before the ego vehicle arrives at the intersection and comparing this to the current position and velocity.

The beliefs represented by the world model may be revised with time due to the arrival of new information. One possibility is to refine an imprecise perception. E.g., we weren't sure what the entity was until it got close when we realised it is a bicycle. Another (non-monotonic) possibility is to correct a misperception. E.g., we thought an entity was a motorcycle until it got close and we realised it is a bicycle.

2.1 Integration of the world model schema with RSS

RSS is intended to be a minimal specification of safe driving behaviour that is refined by AD implementations. Thus we assume that RSS provides a specification in the form of an *RSS basic world model schema* and the schema of any AD implementation conforming to RSS must refine this schema. A concrete link to safety assurance here is that the argument, or ideally, proof, that this refinement is correct is part of the safety case for AD system.

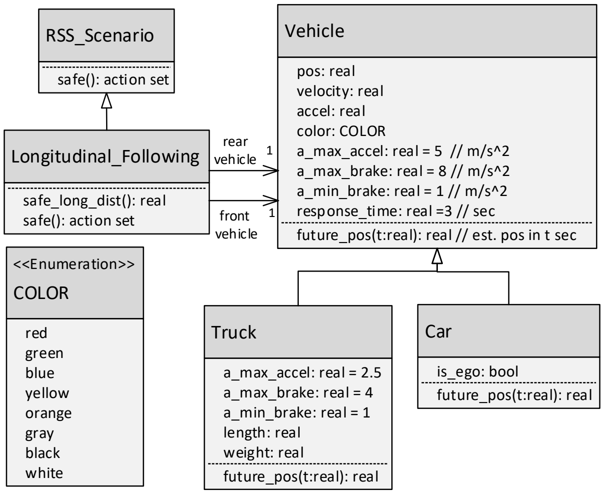

Figure 1: A fragment of an hypothetical RSS compliant world model schema showing classes for the running example visualized as a UML class diagram.

The RSS basic world model schema should provide definitions of base classes (e.g., `Road_user`), and all attributes and parameters required by RSS rules. In addition, it defines the function `Safe` where `Safe`(s) is the set of safe actions for world model s. This can be defined in terms of safe distance to other road users, acceleration, and other criteria. Thus, in a dangerous situation, `Safe`(s) is restricted to actions corresponding to the proper response. In other situations, `Safe`(s) contains the set of actions that will not cause a dangerous situation to be entered due to the ego vehicle.

Fig. 1 shows a fragment of a hypothetical RSS compliant world model schema showing classes for the "safe longitudinal distance, same direction" scenario visualized as a UML class diagram. The class `Vehicle` has been refined here to distinguish two classes that have different parameters used in RSS computations. We assume typical inheritance/override semantics so the parameters of class `Car` are the same as for `Vehicle` but not for `Truck`. Only one scenario class, `Longitudinal_Following`, is shown with attributes (modeled here as UML associations) indicating the front and rear vehicles. Every subclass of `RSS_Scenario` must define function `Safe`, producing the set of safe actions in the current state of the scenario. In the case of the longitudinal following scenario, this corresponds to actions of the rear vehicle and also uses a function `safe_long_dist` to compute d_{min}.

2.2 Misperception with precise world models

To study misperception, we can compare perceived states to true states. Let $s \rightarrow s'$ denote a *perception case* where $s, s' \in \mathcal{S}_{\Sigma}$, s is the correct world model (i.e. ground truth) and s' is what is perceived by the perceptual subsystem. A perception case summarises a possible behavior of the perceptual subsystem in perceiving the world at a given point in time. Perception case $s \rightarrow s'$ is a *correct perception* iff $s' = s$; otherwise, it is a *misperception*.

To illustrate perception cases, assume the correct world model s has a single instance of the Longitudinal_Following scenario from Fig. 1 with the rear vehicle being a Car and the front being a Truck. Examples of misperception cases $s \rightarrow s'$ include: s' classifies the front vehicle as Car, s' has an erroneous velocity value for the rear vehicle, s' is missing the front vehicle, and s' does not contain an instance of Longitudinal_Following.

RSS is designed to be safe when perceptions are correct. Some misperceptions are benign — e.g., misperceiving a gray Vehicle as being white. Others, such as the examples above, can pose a safety risk. We formally define when a misperception can pose a safety risk.

Definition 2.1. Misperception case $s \rightarrow s'$ potentially causes safety risk iff $\mathtt{Safe}(s') \not\subseteq \mathtt{Safe}(s)$.

That is, there is a potential safety risk when the safe actions for the perceived world model include actions that are not safe in the true world model. Note that there is a *definite* safety risk when $\mathtt{Safe}(s') \cap \mathtt{Safe}(s) = \emptyset$; however, even when $\mathtt{Safe}(s') \subset \mathtt{Safe}(s)$, and the misperception is safe, the fact that $\mathtt{Safe}(s')$ is a proper subset means that it restricts the allowable safe actions. This can negatively impact non-safety quality criteria such as progress or comfort.

3 Incorporating uncertainty

The world model represents the current beliefs of the AD system and we assume that we have information about the degree of uncertainty (or conversely, confidence) about these beliefs. This can come from development-time safety assurance analyses of the perceptual subsystem or directly during operation from the perceptual subsystem itself. The latter is a reasonable assumption for ML-based perceptual components, since their output often includes a value that can be interpreted as a measure of confidence. For example, when used for classification, a support vector machine (SVM) outputs a classification as well as the distance to the decision boundary, where being further from the boundary indicates higher confidence. Similarly, the softmax output of a deep neural network (DNN) can be interpreted as a categorical distribution and the probability value associated with the classification decision indicates the confidence. Estimating the uncertainty of neural network predictions is the subject of active research (Malinin and Gales 2018).

A pure Bayesian approach to using this uncertainty information requires propagating a probability distribution over world models into the RSS rules to generate a corresponding distribution over safe actions; however, this approach is known to be intractable for all but simple cases. Instead, we approach the handling of uncertainty by defining *imprecise* world models that represent different levels of uncertainty.

3.1 Imprecise world models

Semantically, an imprecise world model $\bar{s}$ represents a set of precise models $[\bar{s}] = \{s_1, s_2, ...\}$. Let $s \rightarrow_\alpha \bar{s}$ denote an *under-perception* case, where $\bar{s}$ is an imprecise model *perceived with confidence* α when the correct model is s. This is interpreted as saying "the AD system believes that, with probability α, the precise model s is in set $[\bar{s}]$" or, formally,

$$(s \rightarrow_\alpha \bar{s}) \Rightarrow \mathtt{Pr}(s \in [\bar{s}]) = \alpha \qquad (1)$$

Thus, $\bar{s}$ represents the $\alpha \times 100\%$ *credible set* of s (i.e., the Bayesian equivalent of a confidence set). It may be that Eq. 1 depends on certain conditions, such as the perceptual subsystem being *calibrated* (Guo et al. 2017) — i.e., that the uncertainty is correctly aligned with accuracy. We leave the investigation of the conditions on Eq. 1 for future work.

$$\mathtt{s_c} = \{\mathtt{o1} : \mathtt{Car}\{\mathtt{pos} = 0\}, \mathtt{o2} : \mathtt{Car}\{\mathtt{pos} = 30\}\}$$
$$\bar{\mathtt{s}}_1 = \{\mathtt{o1} : \mathtt{Car}\{\mathtt{pos} = 0\}, \mathtt{o2} : \mathtt{Truck}\{\mathtt{pos} = [29, 31]\}\}$$
$$\bar{\mathtt{s}}_2 = \{\mathtt{o1} : \mathtt{Car}\{\mathtt{pos} = 0\}, \mathtt{o2} : \mathtt{Vehicle}\{\mathtt{pos} = [28, 32]\}\}$$

Figure 2: Under-perception examples.

A world model schema as described in Sec. 2 represents precise models but it can be *lifted* to represent imprecise models with confidence α. The class of an entity can be made imprecise by using a super-class while a continuous attribute for which a probability distribution is given by the perceptual system can be made imprecise by using a credible interval as its value. For example, given the schema in Fig. 1, assume the true current world model is $\mathtt{s_c}$ as given in Fig. 2. We omit some attributes for simplicity. Assume the perception system assigns entity o2 probability 0.7 of being Truck and 0.3 of being Car, and its attribute pos is perceived as normally distributed according to $\mathcal{N}(30, 1)$. In this case, using $\alpha = 0.68$, would produce under-perception $\mathtt{s_c} \rightarrow_{0.68} \bar{\mathtt{s}}_1$ where o2 is classified precisely (but, incorrectly) as Truck and the position is confident to an interval of 1 standard deviation. If the higher confidence $\alpha = 0.95$ is required, the under-perception $\mathtt{s_c} \rightarrow_{0.95} \bar{\mathtt{s}}_2$ classifies o2 less precisely as Vehicle and pos is given as a larger interval of 2 standard deviations. Note that $\mathtt{s_c} \notin [\bar{\mathtt{s}}_1]$ but $\mathtt{s_c} \in [\bar{\mathtt{s}}_2]$ due to the less precise world model produced by the higher confidence requirement.

The computations of derived attributes used in RSS rules must be reinterpreted for imprecise values. For example, in scenario Longitudinal_Following, the computation of safe_long_dist() (i.e., d_{min}) in terms of imprecise velocity attributes (i.e., intervals) for the rear and front car should use *interval arithmetic* (now covered by standard IEEE 1788 (Revol 2017)).

The key property that lifting function Safe to imprecise model $\bar{s}$ should have is the following:

$$\mathtt{Safe}(\bar{s}) = \bigcap_{s_i \in \bar{s}} \mathtt{Safe}(s_i) \qquad (2)$$

That is, *a safe action in an imprecise model must be safe for every precise model covered by the imprecise model.* We can argue that Eq. 2 holds for the vehicle classification example discussed above. The imprecise classification as Vehicle (corresponding to $\alpha = 0.95$) represents the set $\{\mathtt{Car}, \mathtt{Truck}\}$ of precise classes. Since the parameters shown in Fig. 1 for class Vehicle are at least as conservative as for its subclasses, any action that RSS considers safe when treating the front vehicle as Vehicle will also be safe if the vehicle is treated as a Car or Truck.

3.2 Using imprecise world models to mitigate misperception

The most important consequence of Eq. 1 and Eq. 2, is that for any under-perception $s \rightarrow_\alpha \bar{s}$,

$$\Pr(\texttt{Safe}(\bar{s}) \subseteq \texttt{Safe}(s)) \geq \alpha \tag{3}$$

Eq. 3 says that *we can make perception as safe as we like and avoid risk associated with misperception in Defn. 2.1 by using under-perception and sufficiently high* α. The reason for this is that Eq. 2 forces the planning/control subsystem to act conservatively and only take actions that are safe for any precise model covered by $\bar{s}$. However, this will not hold if the set $\texttt{Safe}(\bar{s})$ is empty — i.e., there is no safe action in common! This case is not ruled out by Eq. 2 but RSS provides an important safeguard against this, since it is designed so that *there exists a proper response for every combination of unsafe situations*. Thus, regardless of what set of dangerous situations are covered by $\bar{s}$, the set $\texttt{Safe}(\bar{s})$ will always contain proper response actions.

With this, we have shown how the use of under-perception addresses case (1) in Sec. 1.

4 Discussion and Next Steps

Although we have a strong support on the safety of using under-perceptions, there are limitations and impacts to this, and a key part of this research is to explore these and how they may be mitigated. In particular, there is an evident trade-off that strengthening safety may incur a reduction on non-safety quality criteria, such as progress and comfort, since increasing α may cause the set $\texttt{Safe}(\bar{s})$ to shrink and therefore restrict allowable actions. For example, in the classification example discussed above, $\alpha = 0.95$ corresponds to the imprecise classification as `Vehicle`. Since, the parameters for class `Vehicle` are at least as conservative as for its subclasses, any action that RSS considers safe when treating the front vehicle as `Vehicle` will also be safe if the vehicle is treated as `Car`. However, if the front vehicle actually is a truck, then these more conservative parameters will cause d_{min} to be unnecessarily large and may impact progress objectives.

There are two obvious ways to address this trade-off. First, α can be made lower, and in fact, it can vary for different components of the world model schema — but this implies that more risk is incurred and there is lower bound to what is societally acceptable. Second, the perceptual system could be improved so that it exhibits higher confidence in its perceptions and hence the size of the set $[\bar{s}]$ decreases (which causes $\texttt{Safe}(\bar{s})$ to increase). This may be possible, but there are factors that limit it as well, including aleatoric uncertainty. We are interested in investigating what approaches may be available.

A key concern in this research is to make the approach scalable to complex world model schemas and large world models. RSS addresses scalability for planning and control by ensuring that correctness of the rules can be established compositionally by only considering local interactions between road users (i.e., the so-called "star-shaped" argument). A similar compositionality is needed for working with imprecise models using under-perception. One possibility is to decompose the world model schema by treating components such as classes and attributes independently, and the examples in this paper have used this perspective; however, the probability distributions expressing uncertainty may force dependencies. For example, the uncertainty on `Vehicle` position and velocity attributes will be dependent on the state of environmental conditions which are tracked in another part of the world model.

Another aspect of scalability can be addressed by analysing what parts of the world model are *safety-relevant*. When $\texttt{Safe}(\bar{s}) = \texttt{Safe}(s)$, the under-perception is safety *irrelevant* — i.e., whatever, dimensions of variation are captured by $\bar{s}$, they are not relevant to safety decisions in world model s. For example, examining the d_{min} computation on which $\texttt{Safe}(s)$ is based in the longitudinal following scenario shows that it depends only on the velocities of c_r and c_f. This means that under-perception of attributes of other cars outside the scenario (e.g., parked in a driveway), do not affect $\texttt{Safe}(s)$. Identifying the relevant parts of the world model can help focus limited computing resources. This shows an interesting symmetry in the AD architecture: while the perception subsystem contributes the uncertainty information to the world model, the planning and control subsystem contributes the relevance information. Also, note that relevance is specific to quality attribute, so that safety irrelevant aspects of the world model may still be relevant for non-safety quality attributes.

Finally, while not considered in this paper, it is important to explore how the Operational Design Domain (ODD) as well as temporal aspects of world model evolution interact with uncertainty and safety. In an AD system, the possibility of exiting the ODD must be monitored to ensure the safety boundaries of the architecture. Thus, the world model schema must be rich enough to express the ODD conditions. An interesting question to explore is whether under-perception could be used to "soften" the ODD boundaries without compromising safety. For example, if the ODD requires that the AD system operate only in dry weather, and it begins to rain lightly, an indication of increased perceptual uncertainty could be used and a hard ODD exit maneuver could be delayed until the rain intensity exceeds a threshold. Uncertainty about entities can also increase or decrease as the world model evolves and beliefs can be revised. Domain knowledge in the form of behavioural models for different classes in the world model schema could be used to set expectations about entity behaviour, and if these expectations are violated, this could be a basis for increasing uncertainty about it, treating it as "suspicious".

In this paper we have discussed just a few of the topics suggested by this work. Ultimately, we hope this work makes an effective contribution to the important topic of using ML safely in the context of AD systems.

References

Czarnecki, K., and Salay, R. 2018. Towards a framework to manage perceptual uncertainty for safe automated driving.

In *International Conference on Computer Safety, Reliability, and Security*, 439–445. Springer.

Czarnecki, K. 2018a. Operational design domain for automated driving systems: Taxonomy of basic terms. *Waterloo Intelligent Systems Engineering (WISE) Lab, University of Waterloo, Canada*.

Czarnecki, K. 2018b. Operational World Model Ontology for Automated Driving Systems—Part 2: Road Users, Animals, Other Obstacles, and Environmental Conditions. *Waterloo Intelligent Systems Engineering Lab (WISE) Report*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1321–1330. JMLR. org.

Malinin, A., and Gales, M. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 7047–7058.

Revol, N. 2017. Introduction to the IEEE 1788-2015 standard for interval arithmetic. In *International Workshop on Numerical Software Verification*, 14–21. Springer.

Shalev-Shwartz, S.; Shammah, S.; and Shashua, A. 2017. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*.