WIP: *Savvy*: a Trustworthy Autonomous Vehicles Architecture

Ali Shoker, Rehana Yasmin, and Paulo Esteves-Verissimo
Resilient Computing and Cybersecurity Center (RC3),
Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE),
King Abdullah University of Science and Technology (KAUST)
Thuwal 23955-6900, Kingdom of Saudi Arabia
{ali.shoker, rehana.yasmin, paulo.verissimo}@kaust.edu.sa

***Abstract*—The increasing interest in Autonomous Vehicles (AVs) is notable, driven by economic, safety, and performance reasons. Despite the growing adoption of recent AV architectures hinging on the advanced AI models, there is a significant number of fatal incidents. This paper calls for the need to revisit the fundamentals of building safety-critical AV architectures for mainstream adoption of AVs. The key tenets are: (i) finding a balance between intelligence and trustworthiness, considering efficiency and functionality brought in by AI/ML, while prioritizing indispensable safety and security; (ii) developing an advanced architecture that addresses the hard challenge of reconciling the stochastic nature of AI/ML with the determinism of driving control theory. Introducing *Savvy*, a novel AV architecture leveraging the strengths of intelligence and trustworthiness, this paper advocates for a *safety-first* approach by integrating design-time (deterministic) control rules with optimized decisions generated by dynamic ML models, all within *constrained time-safety bounds*. *Savvy* prioritizes early identification of critical obstacles, like recognizing an elephant as an object, ensuring safety takes precedence over optimal recognition just before a collision. This position paper outlines *Savvy*'s motivations and concepts, with ongoing refinements and empirical evaluations in progress.**

I. Introduction

The Autonomous Vehicles (AVs) market is in steady growth at a CAGR of 30%, hitting the market size of 2 Trillion USD by 2030. While the performance motivations of AVs are obvious, e.g., through optimizing the driving experience through situational awareness, the safety benefits are little understood. A recent study by RAND [25] argues that "delaying full deployment of AVs until an extraordinarily high level of safety is achieved in comparison to human drivers could cost hundreds of thousands of lives over many years" [52]. This study seems to call for an unregulated path for the development and deployment of AVs, especially under the later AI/ML-based paradigms, despite accidents in the way. Unfortunately, the increasing global fatal incidents of AVs do not look satisfactory or encouraging, urging the need to revisit the fundamentals of building safety-critical AV architectures and solutions.

Symposium on Vehicles Security and Privacy (VehicleSec) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-7-6
https://dx.doi.org/10.14722/vehiclesec.2024.23058
www.ndss-symposium.org

Our claim is that we need to strike a balance between intelligence and trustworthiness, that is, between the appealing efficiency and rich functionality brought in by AI/ML paradigms, and the indispensable social duty of ensuring safety and security of transportation. Technically, reconciling the AI/ML stochastic nature with the determinism of driving control theory has been a hard challenge, and subject of several research attempts that we report ahead. We argue that the balance must not exclude any of these paradigms. While we adopt the "Safety First for Automated Driving" (SaFAD) [7] in this work, which is also promoted by the European Association of Automotive Suppliers, a practical AV architecture must not ignore efficiency and performance.

Current AV architectures capitalize heavily on the recent advances in AI/ML. Unfortunately, AV architectures in production like *Tesla, Cruise, Waymo*, or *Udacity* have gained bad reputation given the increasing death incidents [8], [15], [47] attributed to the failure of AI/ML systems in particular. The investigations reported in Section II conclude that the majority of these incidents are caused by the tendency to *prioritize performance at the expense of safety*.

We conducted an analysis of real-world AV reported incidents (details discussed in Section II) in order to understand the current phenomena behind AV failures. Our conclusions attribute these failures to two main reasons: (1) Confusion in command and control, and (2) AI-based AVs are optimized for better-precise-than-timely decision processes resulting in All-or-Nothing dilemma. The two reasons are very related: the system is often confused and cannot make a decision because the AI system has not delivered early enough, or never, before the incident. It seems that the AI system often tries to optimize detection, recognition, and planning which often exceeds the available time bounds. This has been raising concerns in the research community for a decade now, which motivated several academic works like Safe-AV [42], Sentinel [10], *KARYON* [5] or E-GAS [30], to address this issue. These works have built on early conceptual and architectural works, e.g., *Timely Computing Base* [33], [49] and *Simplex* [41], [48]. In many cases one could infer that safety-first fall-back mechanisms did not always succeed, for example hand-over, on time, to a module securing fail-operational behavior. Moreover, in the best case when it succeeds to fail-operational, the AV completely sacrifices the power and cost of the AI system.

The above observations inspired us to explore an innovative mechanism, *Time-aware Predictive Quality Degradation,*

TABLE I. A SAMPLE OF INVESTIGATED VEHICLE INCIDENTS WITH DIFFERENT LEVELS OF AUTONOMY, REFERRED TO COMMON POTENTIAL CAUSES.

ID	Incident	Description	Potential Causes
I1	Uber Volvo XC90 (Arizona, 2018) [37], [42]	A modified Volvo XC90 struck and killed a pedestrian walking a bike and crossing a road at night. The vehicle was equipped with a LIDAR unit, forward facing and side facing cameras, radars and Uber's developmental AV software. The car did not brake or attempt to slow down to avoid the collision. Emergency braking and driver monitoring were disabled. NTSB report states that the vehicle's sensors detected the pedestrian 6 seconds before the accident; initially detecting her as an unknown object, then as a vehicle, and finally as a bicycle; 1.3 seconds before impact, the vehicle's software determined that a braking action was required.	(1) Emergency braking and driver monitoring were disabled. (2) Braking decision has been made 4.7 seconds after first detection.
I2	Acura MDX (Newfoundland, 2018) [42]	Acura MDX equipped with Acura's lane-keep assistance attempt to veer off of its lane and sometimes into oncoming traffic. This was noticed after replacing vehicle's windshield.	The camera, a crucial component for lane-keep feature, was not calibrated after windshield replacement.
I3	Tesla Model X (California, 2017) [42]	Tesla Model X (without LIDAR) while using Autopilot feature (cruise control and autosteer lane-keep) crashed into a damaged crash highway attenuator and fatally wounded its driver. At three seconds prior to the crash and up to the time of impact with the crash attenuator, the Tesla's speed increased, with no precrash braking or evasive steering movement detected.	(1) Attenuator has not been detected yet. (2) No safety short circuit caused the car to slow-down or stop. (3) Sensors are not used effectively.
I4	Tesla Model S (California, 2018) [42]	Tesla Model S (with Autopilot enabled) was travelling on a freeway crashed into a stopped fire truck. The Tesla was following another vehicle that swerved out of the lane to avoid the stopped fire truck, while the Tesla sped up instead, and crashed into the truck.	(1) Cruise Control failed to detect the stationary truck on time. (2) Cruise Control detected but ignored the truck.
I5	Tesla Model S (Florida, 2016) [42]	Tesla Model S with Autopilot engaged struck and passed beneath a coming tractor trailer that was making a left turn infront of the Tesla from the westbound lanes of the highway across the two eastbound travel lanes. NTSB reported that the Tesla's automated vehicle control system did not identify the truck crossing the car's path or recognize the impending crash; consequently, the Autopilot system did not reduce the car's velocity, the forward collision warning system did not provide an alert, and the automatic emergency braking did not activate. Tesla commented that the camera failed to detect the truck due to "white colour against a brightly lit sky" and a "high ride height", and that the radar filtered out the truck as an overhead road sign to prevent false braking.	(1) Truck crossing has not been detected on time. (2) Safety circuit has not engaged the emergency braking. (3) Sensors are not utilized effectively. (4) Detection has been ignored.
I6	Tesla Model S (China, 2016) [42]	Tesla Model S crashed into a slow moving (or parked) street sweeper and killed its driver. The police concluded that the neither the driver nor the vehicle had attempted any braking or collision avoidance manoeuvres. Tesla was equipped with a single forward facing radar, a single forward facing camera and a set of 12 ultrasonic sensors. While the camera used DNN recognition models over MobileEye's EyeQ3 computing platform, the system required agreement between both the camera and the radar before any action was taken.	(1) Camera system failed to detect the sweeper on time; (2) Camera and Radar both failed. (3) Detection of Radar alone has been ignored.
I7	GM Cruise (San Francisco, 2022) [23]	Cruise vehicle operating in autonomous mode made a left turn in front of an oncoming Toyota Prius and preformed hard brake at an intersection. NHTSA reported that the Cruise's ADAS could make "unprotected left, cause ADAS to incorrectly predict another vehicle's path or be insufficiently reactive to the sudden path change of a road user." Cruise said the software had to decide between two different risk scenarios: hard brake or collide before the oncoming vehicle's sudden change of direction".	(1) ADAS cannot predict path on time. (2) Pre-defined decisions (turn left) is not always reasonable.

TPQD: using dynamic ML models that can be tuned to provide *either* richer *or* faster outputs based on the available safety-critical time bounds. That is, to tune ML models to enforce different richness levels or predictive quality dictated by the available time-safety bounds. This allows to leverage the *best outcome an AI system can deliver within a given time interval*. We show in this paper that, in many scenarios, AI models seem to be over-optimized to give rich predictive details, while basic details can be good-enough for situations where time is paramount. Based on these concepts, we propose *Savvy*, the preliminary design of a new AV architecture that stands as a sweet spot between performance and safety. *Savvy* ensures the safety-first principle by combining the *time-elasticity* and *time-safety* [49] enabled by the TPQD and the fail-operational control mechanisms. This is possible using Dynamic AI models that can be predictably tuned to deliver before the safety-critical time expires. We are exploring Dynamic Neural Networks that allow for model deformation using depth and width adjustment [4], [21], [31], [46], [51] (early exiting, skipping, pruning, etc.), choosing the adequate protocol using Neural Architecture Search [57], [59], or parameter (Weights, Space, or Channel) adjustments [6], [19], [29], [50], [53] at inference time. We are currently implementing a proof of concept of *Savvy* and driving an empirical evaluation of TPQD.

The rest of the paper is organized as follows. Section II presents the motivations behind *Savvy* through analyzing AV incident investigations. Section III introduces *Savvy* architecture, while Section IV analyzes the pros and cons of *Savvy*. Section V discusses the related work and the paper is concluded in Section VI.

II. THE CASE FOR *Savvy*

This work is driven by our concerns about the hundreds of vehicle incidents [13], [40], [42] related to autonomous driving features at any SAE level of driving automation, including driving assistance, emergency brake, auto-steering, cruise controls, etc. Analyzing dozens of these reported incidents (mostly under investigations), we observed that despite their different circumstances, the majority can be attributed to a common set of potential causes conveyed in Table I, which stands as a briefing of seven credible well-reported and investigated incidents. We are inspired by these observations to explain the design rationale behind *Savvy* (discussed in Section III):

A. Confusion in Command and Control

We observed that many incidents occur because of the confusion in "command and control" in the autonomous driving (AD) system. In many cases, the control is either retained by the data plane, i.e., ML-based system, or "lost" because of some confusion in the system or handover. While an ML-based system does have promising potential, even for safety features, it cannot fully retain AD control because (1) it is a probabilistic solution, and (2) it does not oversee the entire vehicle state. For instance, incidents I3 through I7 in Table I convey different AD failures in detection or prediction. Worse, in some cases like I1, I3, and I5, the vehicle ignored the sensing alerts and has

not made any slow-down or braking reactions; either because the AD system could not make that decision or because a last moment handover was being done. The AD system should have at least handed over the control to some safety-circuit in such a hazardous situation. On the other hand, many incidents like I1, I2, I4, and I6 are referred to the lack of overseeing or handling the vehicle posture, e.g., features disabled, broken or non-calibrated sensors/actuators, or radars sensing ignored. The vehicle state should be retained by a reliable system that can oversee the entire vehicle state as well as *take over, not get handed over* the control in such critical situations where ML fails to respond on time.

B. Better-precise-than-timely decision processes

In all incidents of Table I, the ML-based AD failed to recognize an obstacle or predict a plan or a maneuver on time. The term "on time" here is key since no reports mentioned the AD returning an "invalid" or "indeterminate" classification or prediction, but rather AD *has not delivered early enough before* the incident. Notice that in I1, for example, the AD system recognized an unknown object 6 seconds before the incident and took 4.7 seconds to determine a brake is required. I3 and I7 show that an obstacle has been detected, but the system was not able to make a correct prediction or decision on time. Incidents I3, I5, and I6 show that some sensors have not been considered in decision making, maybe for some optimization (avoiding false-negatives) or because the system could not make a timely sensor fusion.

Our hypothesis is that *ML-based solutions are over optimized for better-precise-than-timely decision processes resulting in All-or-Nothing regardless of the delivery time*. While this can be questionable, it inspired us to study the effectiveness of accepting some *predictive quality degradation in ML inference in favor of timeliness* to guarantee safety. This helps in enforcing decision making timeouts, e.g., by calibrating dynamic ML algorithms [4], [21], [31], [46], [51] (e.g., deepness or parameters) to deliver before timeouts as long as the output is helpful, though not optimal.

To exemplify, we convey different scenarios presented in Table II. The table demonstrates that from a single event, there could be different rich/poor sensing and planning levels upon which *safety-critical* "possible actions" can be made the earliest possible. In the Obstacle Avoidance scenario for instance, there are many sensing levels, either because of ML model deepness or considered sensors fusion, that could help making constructive decisions within some known time windows. Depending on the time availability (e.g., before hitting an obstacle), a used ML model is calibrated (or another model is fetched) to give richer details. Note that the *safety-critical* possible actions in the table are not our recommendations, but rather used to explain the concept. (This is worth another study out of the scope of this paper.) This motivated us to designing *Savvy*, a new architecture with this concept as we show next.

TABLE II. DIFFERENT SCENARIOS SHOWING THAT SOME TIME-AWARE PREDICTIVE QUALITY DEGRADATION CAN STILL BE HELPFUL IN DECISION MAKING WITHIN DIFFERENT TIME BOUNDS. (HINT: INCREASING L REFERS TO RICHER ML OUTPUT DETAILS, INCURRING MORE DELAYS.)

Level	Sensing	Possible Actions
	1. OBSTACLE AVOIDANCE	
L1	An object detected at safety distance	brake; beep
L2	Non obstructive shaped (flat, small, short) object detected	continue
L3	Non obstructive material object detected (rubber, herb plant, snow)	continue slowly
L4	Obstructive avoidable object detected	beep; steer away
L5	Obstructive unavoidable material object detected	brake; beep
L6	Obstructive mobile object detected (auto, animal)	brake; give way; continue later
L7	Obstructive rational object (human) detected	brake; stop; continue later
	2. INTERSECTION CROSSING	
L1	No cooperative sensing	brake
L2	Cooperative sensing (e.g, RSU) short distance	brake
L3	Cooperative sensing (e.g, RSU) long distance	continue
L4	Cooperative active sensing	agreement
	3. OVERTAKING	
L1	No cooperative sensing	continue
L2	Cooperative sensing (e.g, RSU) short distance	slow down
L3	Cooperative sensing (e.g, RSU) long distance	overtake
L4	Cooperative active sensing	agreement
	4. CRASH AVOIDANCE	
L1	No cooperative sensing	default (brake)
L2	Cooperative sensing (e.g, RSU) short front distance	stop
L3	Cooperative sensing (e.g, RSU) long front distance	slow down
L4	Cooperative sensing (e.g, RSU) short front and back distance	maneuver
L5	Cooperative active sensing	agreement

III. TRUSTWORTHY AUTONOMOUS VEHICLES ARCHITECTURE

To address the above challenges, we propose a new preliminary AV architecture that we introduce next, while we leave empirical evaluation to future work. However, we first introduce the design decisions behind *Savvy*.

A. Design Rationales

1) Time-aware predictive quality degradation: We bridge the All-or-Nothing gap of the AI system with time-aware predictive quality degradation (TPQD). TPQD specifies that the AI system delivery should be maximized within safety time bounds even if at degraded quality. This can leverage the tuning properties of tunable AI models like DNNs [4], [21], [31], [46], [51]. The intuition is to reduce the likelihood of hitting the safety time bounds and consequently fail-operational because of aiming at high predictive quality and rich recognition. For instance, Savvy enables the safe identification of an elephant as an obstacle object at the earliest possible, instead of classifying it optimally as an elephant when it is too late; and allows to optimally identify a tunnel as is when time permits, rather than being conservative (and maybe slow-down) if otherwise classified as an obstacle object. This decision allows making use of the AI system capabilities as much as possible without violating the safety bounds.

2) Safety-first supervisory control: Safety in an autonomous vehicle is paramount. To be able to make decisions without confusions, we enforce some centralized control where processes are coordinated by a *Safety-critical Supervisory Control System* (SSCS). This inherits the safety-first principles of recent architectures [5], [7], [10], [30], [41], [42], but

importantly controls and monitors the time safety bounds across processes, including the AI system. The SSCS can benefit from *AI-based Delivery Time Estimation* (henceforth *TED*) models to estimate the delivery time of AI processes. In our experience, the inference time of pre-trained DNN models is very predictable. The challenge part is launching the time of *triggering events*. For this, while we maximize the use of the available sensing and actuating capabilities, we require a quick *bird's eye sensing* based on which the SSCS scheduling process is sparked. In this vein, we encourage introducing more sensing technologies for that very purpose.

B. Savvy Architecture

To implement the aforementioned design decisions, we propose the *Savvy* architecture, comprising two main components: a safety-critical supervisory control system referred to as the *Safety-Critical Control* (SCC) system, and an AI-based system that may include multiple *Time-Sensitive Intelligent Modules* (TSIMs) based on the perception/planning model of AV. Fig. 1 depicts the *Savvy* architecture with reference to the well-known Sense-Plan-Act (SPA) model [7], [10], [42].

The SCC assumes full supervisory control of the system including time scheduling of AI-based TSIMs tasks. The SCC system employs a time-sensitive *Task Scheduler* (TS) that generates the time bounds for the driving related tasks, ensuring the safety-critical timeliness. For each driving related event, TS defines the task schedules across the different TSIMs and set the safety timers accordingly. In particular, TS defines two time bounds that represent the time interval [*Time to Hazard (TTH), Time to Event (TTE)*]. The TTE, e.g., Time to Curve, Time to Overtake, etc., is the time assigned to each TSIM to deliver. TTH, on the other hand, defines the safety-critical time by which the whole driving task should complete, guaranteeing the safety.

A TSIM processes driving tasks, e.g., detection, respecting the time bounds. Each TSIM is further composed of the Static submodule (SMod) and the Dynamic submodule (DMod). DMod leverages the dynamic AI models capabilities that are tuned, for instance, using Bounded AI (such as Neurosymbolic NNs [9], [14], [17] and Physics Informed Neural Networks (PINN) [1], [26], [34], [56] models), to deliver before the TTE expires. Tuning is done via a time prediction system that learns and estimates the calibration parameters of AI models to be used in order to deliver before the TTE expires. The TSIMs are self-contained modules that can be used in any architecture; for instance, Fig. 1 demonstrates the well-known Sense-Plan-Act model using three TSIMs for the three modules of Sense, Plan, and Act.

The workflow starts by a sensing trigger issued by a *preliminary-sensing module* that is tailored for quick bird's eye detection. This may leverage any available sensors to send a heads-up to the SCC, activating a new driving task process. The preliminary-sensing feeds the SCC with initial time boundaries based on which the SCC can schedule (using time prediction models) fine-grained tasks over the TSIMs, e.g., in this case over the three TSIMs of SPA modules. Different TSIMs can interact as necessary to do more sensing, fusion, perception, planning, etc. According to the proposed time limits given, the TSIMs tune the Dynamic AI models

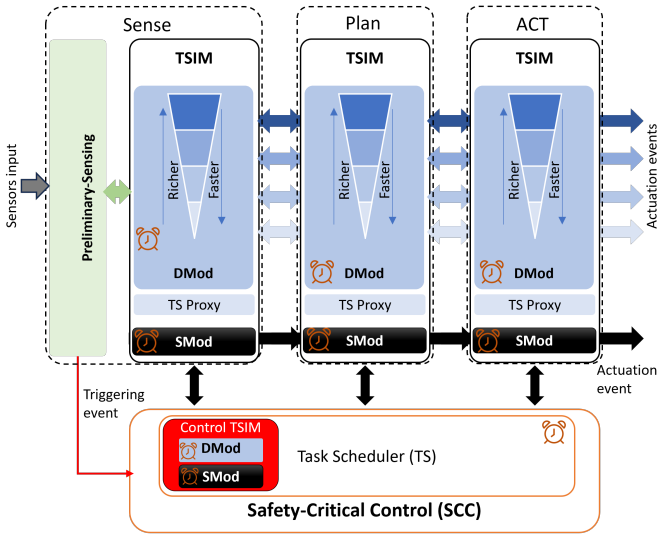

Fig. 1. *Savvy* architecture demonstrated on the Sense-Plan-Act model.

to meet the delivery time. This may sometimes reduce the prediction quality, but should deliver useful insights, e.g., frames of animals versus animals, shadows versus liquids on road, etc. The TTE is the expected time by which all the DMods should deliver and execute in order to take advantage of the AI capabilities. In the event where the TSIM AI execution, that is DMod processing, hits the lower time bound, the SMod is triggered by firing the safety timer TTH. The TTH defines the deadline to execute all the safety executions of SMods to guarantee global fail-safe.

To estimate TTH and TTE, TS again uses a special Control TSIM whose SMod is defined at design time, and DMod is made highly accurate using real accurate formulas or more Bounded AI models. The TS distributes the TTH and TEE over all TSIMs following some policy: statically, e.g., on evenly basis, or dynamically, in a similar to DMods. At any time, before scheduling or during running TSIM, an expired timer will immediately launch a *safety-critical* action controlled by the SCC (i.e., TSIM's AI never assumes control).

The SCC system plays the role of Pub/Sub broker where sensors push their readings to, and actuators take order from. Upon the receipt of input from that set of sensors, that *triggers the event*, the SCC system triggers the corresponding set of actuators before the safety-critical time TTH, if better opportunistic decisions are not expected within the TTE. With this, the *safety-critical* action is always guaranteed to be triggered and completed before a critical event or time TTH, beyond which the vehicle would be at risk.

IV. Discussion

A. Feasibility

Savvy is a practical architecture as it is aligned with the recent trend of centralized or semi-centralized vehicle E/E architectures where a powerful central Electronic Control Unit (ECU) of cluster of ECUs handle the heavyweight processing. This is however assumed that this central brain is dependable and secure, for instance, it could be a replicated state machine over many ECUs or COTS in System on Chip (SoC).

Savvy architecture may impose some additional computing and AI training overheads as it mandates the separation of control plane and data plane, and it also introduces additional supervisory *safety-critical control* layer to the AV architecture. The run-time model-tuning and time-aware predictive quality degradation may also add some AI training costs. However, this entire overhead comes at the expense of prioritizing the safety of precious lives, aligning with the principle of 'Safety First for Automated Driving'. A Proof of Concept for Savvy is currently in the implementation phase. This will facilitate an empirical evaluation of the proposed trustworthy autonomous vehicle architecture, *Savvy*.

B. Safety versus Convenience

The 'all or nothing' strategy, employed by other AV architectures, aims at achieving an optimized driving performance, as too many safety triggers may frequently slow down the vehicle, resulting in an unpleasant passenger experience. However, this strategy also presents a significant risk to human lives and road safety due to its inability to deliver before timeouts in several cases discussed in detail in Section II. *Savvy*, on the other hand, leverages the dynamic AI models capabilities before the TTE expires just like other AV architectures. However, the fail-safe mode is only triggered if the TTE timer expires that is the safety-critical time beyond which the road-safety is at a higher risk. The predefined tasks of SMode may not always be the most efficient solution, compromising the driving performance to some extent, however, they will reduce the safety incidents which otherwise happen in case of 'all or nothing' strategy. *Savvy* stands as a quality degradation method to exploit the best possible of the AI usage; and therefore, avoids the *all-or-nothing* dilemma. For instance, it enables the safe identification of an elephant as an obstacle the earliest possible, as opposed to attempting optimal classification as an elephant when it might be too late. It also allows for optimal classification of an elephant as is when time permits, avoiding unnecessary conservatism.

C. Safety-Critical predefined tasks

Due to dynamic nature of *Savvy*, the *safety-critical* predefined tasks will be defined as the optimized ones considering the road circumstances. These predefined tasks will be based on human driver's posture, mental and psychological state, driving skills, as well as the real-time situation awareness. For instance, the decision could be to continue driving and apply a brake when a very little distance (time) is left between the vehicle and the object, if the AI model fails to deliver. However, as the sudden brake at a busy road may likely be risky, the decision could be to gradually slow down the vehicle speed, provided that the distance between the vehicle and the object is known (like in *Savvy*), and apply a brake only when TTE expires; enabling the vehicle to apply a brake just before the safety-critical time beyond which a safety incident may happen, if the AI model fails to deliver.

We exemplify the idea using two cases conveyed in Table I. In case I1 of Uber Volvo XC90 incident, the object was detected 6 seconds before the accident whereas 4.7 seconds were taken in object classification, and braking decision was made just 1.3s before the accident which could not be implemented in this short time of 1.3s. Once the brake process is initiated and the vehicle's braking system comes into play, the vehicle needs some time to come to a complete stop after applying the brakes depending on various factors such as vehicle's initial speed, braking system efficiency, road conditions, etc. The gap of 6 seconds was already too short and if the SCC had been activated, braking or speed-reducing decisions would have been initiated earlier, it would have been feasible to complete the braking process before the incident. Similarly, in case I6 of Tesla Model S incident, the system required an agreement between both the camera and the radar before any action could be taken. The detection of the Radar alone had been ignored. It's believed that the possible cause for this accident might have been the camera and its ML algorithm failing to detect the object. If based on the Radar input alone, the SCC had been activated, it might have avoided the causality. As mentioned earlier, the predefined tasks necessitate further investigation, which is beyond the scope of this paper.

The proposed architecture may not solve all the incidents faced by AVs, for instance, if AV fails due to the inability of sensors to detect an object in time. However, it may handle several others, some of them discussed above, reducing the number of such incidents. Moreover, we also emphasize the need to introduce new sensors capable of detecting objects more easily and at the earliest possible moment.

V. Related Works

A. Safe AV Architectures

Safety-first principles have been used in early architectures of automotive safety-critical systems [41], [54]. The Simplex architecture includes a high-assurance and a (complex) high-performance system that run in parallel. The latter controls the system as long as safety is not violated, in which case the former can take control using a decision logic supervisory circuit. The implementation of the supervisory circuit to make a taking control decision is however complex as explained in [3], due to the tradeoffs between safety and performance. Generalized concepts have been used in the E-GAS standardized architecture EGas [54] and [48], while running multiple levels of diagnostic monitoring and redundancy. Safe-AV [42], a more recent architecture, combines the prior solution in more redundancy levels while also supported ML-based AV. Similar to *Savvy*, these architectures make a clear separation between performance and safety planes, however without addressing ML-based systems or quality/service degradation.

Service level degradation, on the other hand, has been proposed in KARYON [5] to ensure timeliness and switch to "hard-coded" safety kernel. We inspire from this work to propose degradation with ML inference, which has not been addressed in [5]. A recent monolithic architecture, Sentinel [10], has been proposed to cover the aforementioned safety-first concepts with ML degradation. However, degradation in Sentinel is more like cross-validation, as it makes use of a combination of parallel inaccurate predictions (e.g., 60% accuracy) to consolidate a decision. This is far from safe compared to *Savvy*'s degradation technique that uses a degraded classification problem whose accuracy is high. For instance, Sentinel may recognize an elephant with 60% accuracy, while *Savvy* recognizes it as obstructing object with 95%. The latter is a credible accuracy to make an informed

decision, although not optimal, while Sentinel's decision is highly risky.

Unfortunately, we do not discuss commercial architectures like Tesla, Waymo, Cruise, and Huawei-backed AITO M7 as these are not publicly available.

B. AI/ML for AV Architectures

The use of AI/ML in AVs is prominent in recent literature [11]–[13], [16], [32], [39], [43], [52]. In general, most follow the **Sense-Plan-Act design** [7], running several AI/ML architectures and models. The Sense part is focused on sensor data processing including detection, recognition, perception, and localization. It makes use of *Deep Learning* for object detection and classification problems [2], [28] being able learn new features without handcrafted features. In particular, *Convolutional Neural Networks* (CNNs) have shown to be promising for lane and vehicle detection [22]. Plan also uses ML models for prediction and planning. Prediction is essential to guess and evaluate the expected future (e.g., trajectory or behaviour) of the vehicle considering its dynamic surrounding. *Recurrent Neural Networks* [38], [45] (RNNs) are essential to this class of problem, especially *Long-Short Term Memory* (LSTM) networks [20], used to integrate past the present information for end-to-end scene labeling systems. Recent models like Reinforcement Learning [44], [45], mixed with DL can achieve human-level control in [18], [28], and Attention models [35], [55] are being used to improve information filtering. This is mainly useful to focus on the relevant part to "attend" in highly dimensional data, e.g., camera images.

Our work in progress make uses of these techniques and models with two main differences: First, we recommend those models, e.g., **Dynamic Neural Networks** (DNN), that are easily tunable at inference time to be able to leverage the full power of AI/ML despite time limits. Many of the models discussed in this section lie in the category and support dynamic features like: model deformation using depth and width adjustment [21], [31], [46], [51] (early exiting, skipping, pruning, etc.), choosing the adequate protocol using Neural Architecture Search [57], [59] (NAS), or parameter (Weights, Space, or Channel) adjustments [6], [19], [29], [50], [53] at inference time. Second, for more accurate recognition and prediction to improve safety, our work encourages more research and use of what we call **Bounded AI** (BAI) prediction models that include factual pre-trained or symbolic models to guide the training of the main processing models. Recent methods like Neurosymbolic NNs [9], [14], [17], Physics Informed Neural Networks (PINN) [1], [26], [34], Constrained or Conservative PINNs [24], [58], Finite Basis PINN [36], Variational PINN [27] are believed to have close to 100% accuracy in some contexts.

VI. Conclusion

AI is proving to be widely useful especially for non-critical applications. AVs are however more challenging being safety-critical and often time-critical. Reality shows that AI can be useful to improve safe driving compared to humans; however, our analysis shows that AI is not reliable per se to take control of the AV. In particular, due to time-criticality of AV tasks, the architects are either in the conservative camp, and tend to refuge to fail-operational mode often, or in the optimistic camp where performance and user convenience are prioritized at the cost of safety. This paper presents a new AV architecture *Savvy* following a new approach, we call Time-aware predictive quality degradation (TPQD), to combine the two advantages without violating the other. *Savvy* leverages the Dynamic NN properties through tuning them at inference time given the available safety time boundaries. This leads to a trustworthy AV where even for limited time windows, the AI power is being exploited, thus avoiding the *all-or-nothing dilemma*. We are currently implementing the architecture and exploring the DNN models to evaluate empirically.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[3] S. Bak, T. T. Johnson, M. Caccamo, and L. Sha, "Real-time reachability for verified simplex design," in *2014 IEEE Real-Time Systems Symposium*, 2014, pp. 138–148.

[4] S. Cai, Y. Shu, and W. Wang, "Dynamic routing networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3588–3597.

[5] A. Casimiro, J. Kaiser, E. M. Schiller, P. Costa, J. Parizi, R. Johansson, and R. Librino, "The karyon project: Predictable and safe coordination in cooperative vehicular systems," in *2013 43rd Annual IEEE/IFIP Conference on Dependable Systems and Networks Workshop (DSN-W)*. IEEE, 2013, pp. 1–12.

[6] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 030–11 039.

[7] CLEPA, "Safety first for automated driving," *European Association of Automotive Suppliers*, 2019.

[8] R. David Shepardson, "Gm cruise unit suspends all driverless operations after california ban," 2023, last accessed Dec 2023. [Online]. Available: https://www.reuters.com/business/autos-transportation/us-auto-safety-agency-investigating-two-new-gm-cruise-crash-reports-2023-10-26/

[9] L. De Raedt, S. Dumančić, R. Manhaeve, and G. Marra, "From statistical relational to neuro-symbolic artificial intelligence," *arXiv preprint arXiv:2003.08316*, 2020.

[10] S. Deevy, "Sentinel: A software architecture for safe artificial intelligence in autonomous vehicles," Ph.D. dissertation, 2019.

[11] A. Desai, S. Ghosh, S. A. Seshia, N. Shankar, and A. Tiwari, "Soter: a runtime assurance framework for programming safe robotics systems," in *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2019, pp. 138–150.

[12] K. P. Divakarla, A. Emadi, and S. Razavi, "A cognitive advanced driver assistance systems architecture for autonomous-capable electrified vehicles," *IEEE Transactions on Transportation Electrification*, vol. 5, no. 1, pp. 48–58, 2019.

[13] M. R. Endsley, "Autonomous driving systems: A preliminary naturalistic study of the tesla model s," *Journal of Cognitive Engineering and Decision Making*, vol. 11, no. 3, pp. 225–238, 2017.

[14] J. Ferlez, M. Elnaggar, Y. Shoukry, and C. Fleming, "Shieldnn: A provably safe nn filter for unsafe nn controllers," *arXiv preprint arXiv:2006.09564*, 2020.

[15] I. I. for Highway Safety (IIHS), "Driver death rates by make and model," 2023, last accessed 6 July 2023. [Online]. Available: https://www.iihs.org/ratings/driver-death-rates-by-make-and-model

[16] L. Fridman, L. Ding, B. Jenik, and B. Reimer, "Arguing machines: Human supervision of black box ai systems that make life-critical decisions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[17] A. d. Garcez and L. C. Lamb, "Neurosymbolic ai: The 3 rd wave," *Artificial Intelligence Review*, pp. 1–20, 2023.

[18] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[19] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] G. Huang, D. Chen, T. Li, F. Wu, L. Van Der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," *arXiv preprint arXiv:1703.09844*, 2017.

[22] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue *et al.*, "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv:1504.01716*, 2015.

[23] Incidentdatabase.ai, "Incident 293," 2022, last accessed Dec 2023. [Online]. Available: https://incidentdatabase.ai/cite/293/

[24] A. D. Jagtap, E. Kharazmi, and G. E. Karniadakis, "Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems," *Computer Methods in Applied Mechanics and Engineering*, vol. 365, p. 113028, 2020.

[25] N. Kalra and D. G. Groves, *The enemy of good: Estimating the cost of waiting for nearly perfect automated vehicles*. Rand Corporation, 2017.

[26] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[27] E. Kharazmi, Z. Zhang, and G. E. Karniadakis, "hp-vpinns: Variational physics-informed neural networks with domain decomposition," *Computer Methods in Applied Mechanics and Engineering*, vol. 374, p. 113547, 2021.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[29] C. Li, G. Wang, B. Wang, X. Liang, Z. Li, and X. Chang, "Dynamic slimmable network," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 8607–8617.

[30] M. Li and L. Eckstein, "Fail-operational steer-by-wire system for autonomous vehicles," in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, 2019, pp. 1–6.

[31] J. Lin, Y. Rao, J. Lu, and J. Zhou, "Runtime neural pruning," *Advances in neural information processing systems*, vol. 30, 2017.

[32] J. C. Linares, A. Barrientos, and E. M. Márquez, "Hybrid bio-inspired architecture for walking robots through central pattern generators using open source fpgas," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7071–7076.

[33] P. Martins, P. Sousa, A. Casimiro, and P. Verissimo, "A new programming model for dependable adaptive real-time applications," *IEEE Distributed Systems Online, vol. 6, no. 5, 2005.*, May 2005. [Online]. Available: http://www.navigators.di.fc.ul.pt/archive/papers/o5001.pdf

[34] X. Meng, Z. Li, D. Zhang, and G. E. Karniadakis, "Ppinn: Parareal physics-informed neural network for time-dependent pdes," *Computer Methods in Applied Mechanics and Engineering*, vol. 370, p. 113250, 2020.

[35] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[36] B. Moseley, A. Markham, and T. Nissen-Meyer, "Finite basis physics-informed neural networks (fbpinns): a scalable domain decomposition approach for solving differential equations," *arXiv preprint arXiv:2107.07871*, 2021.

[37] N. T. S. B. (NTSB), "Accident Report NTSB/HAR-19/03," 2019, last accessed Dec 2023. [Online]. Available: https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf

[38] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *International conference on machine learning*. PMLR, 2014, pp. 82–90.

[39] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *arXiv preprint arXiv:1704.02532*, 2017.

[40] E. Schwalb, "Analysis of safety of the intended use (sotif)," 2019.

[41] L. Sha *et al.*, "Using simplicity to control complexity."

[42] S. A. Shah, "Safe-av: A fault tolerant safety architecture for autonomous vehicles," Ph.D. dissertation, 2019.

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

[44] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, pp. 9–44, 1988.

[45] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[46] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2464–2469.

[47] Tesladeaths.com, "Tesla deaths," 2023, last accessed 6 July 2023. [Online]. Available: https://www.tesladeaths.com/

[48] M. Törngren, X. Zhang, N. Mohan, M. Becker, L. Svensson, X. Tao, D.-J. Chen, and J. Westman, "Architecting safety supervisors for high levels of automated driving," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 1721–1728.

[49] P. Verissimo and A. Casimiro, "The timely computing base model and architecture," *IEEE Transactions on Computers, vol. 51, n. 8, Aug 2002*, Aug. 2002. [Online]. Available: http://www.navigators.di.fc.ul.pt/software/tcb/papers/TCBjorn.htm

[50] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2017.

[51] X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "Skipnet: Learning dynamic routing in convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 409–424.

[52] N. Webb, D. Smith, C. Ludwick, T. Victor, Q. Hommes, F. Favaro, G. Ivanov, and T. Daniel, "Waymo's safety methodologies and safety readiness determinations," *arXiv preprint arXiv:2011.00054*, 2020.

[53] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[54] E. Workgroup, "Standardized e-gas monitoring concept for gasoline and diesel engine control units," *Version*, vol. 5, p. 38, 2013.

[55] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[56] Y. Yang and P. Perdikaris, "Adversarial uncertainty quantification in physics-informed neural networks," *Journal of Computational Physics*, vol. 394, pp. 136–152, 2019.

[57] J. Yu, P. Jin, H. Liu, G. Bender, P.-J. Kindermans, M. Tan, T. Huang, X. Song, R. Pang, and Q. Le, "Bignas: Scaling up neural architecture search with big single-stage models," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 702–717.

[58] Y. Zhu, N. Zabaras, P.-S. Koutsourelakis, and P. Perdikaris, "Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data," *Journal of Computational Physics*, vol. 394, pp. 56–81, 2019.

[59] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.