Meets Expectations: System Health Analysis and Prognosis for Embedded and Cyber-physical AI

Michael Borth[1], Christian Tiemann[1] and Leonardo Barbini[2]

[1]*TNO, Helmond and* [2]*TNO-ESI, Eindhoven, The Netherlands*
Michael.Borth@tno.nl | Christian.Tiemann@tno.nl | Leonardo.Barbini@tno.nl

ABSTRACT

The health management of AI-based systems differs strongly from that of traditional systems, forcing a rethinking and partial reinvention of established techniques: The AI inside provides the capacity to adapt to an operational context – and that leads to a variability in system behaviors that renders conventional health and performance indicators possibly obsolete. Also, even core AI functionality, like perception, can hardly be assessed without complicated reasoning about whether a lack of performance is circumstantial or an actual system health issue. For these reasons, we introduce a system health monitoring methodology that checks health and key performance indicators against expectations while factoring in mitigating circumstances, like environmental effects. This methodology, which is based on probabilistic reasoning, allows the detection of system health degradation and root-cause analytics and is used by us to ensure the operational fitness of safety-critical systems, i.e., Automated Vehicles. As this domain is subject to temporal changes like seasons that impact a system's performance more than many developing health issues, we combine health monitors with domain monitors and drift detection. Overall, this provides probabilistic health management that looks at expectations, sets of observations, their distributions and their dynamics to determine whether an embedded AI is still fit for its purpose, whether the cyber-physical system embodying it continues to meet the AI's operational requirements, and whether observations indicate a health or fitness trend that will result in a lack of safety. A key aspect of this novel approach to reasoning about system health is that it addresses unique properties of AI-based systems: It works with the hit-or-miss behavior of AI that occasionally fails even on seemingly comparable inputs, and it can investigate adaptive processes by looking at the health of information flows that define the decision-making of AI-based systems.

1. EMBEDDED AI CHALLENGES HEALTH MANAGEMENT

Recent years saw the roll-out of systems that are defined by AI in central aspects of their interactions with the real world. These AI-based systems are referred to as embedded AI systems or cyber-physical AI, depending on the author's preferences and their gestalt, e.g., whether a connection to computer networks is essential. What they have in common is that their system health and thus their ability to perform efficiently, effectively, and safely is defined not only by the absence of failures and faults but also by their fitness – understood as a fit between their AI functionality and the task at hand as well as their operational environment, and the fulfillment of the AI's requirements towards the physical system. Such an extended understanding of system health was introduced by us already in (Borth et al. 2024), where we stress the importance of an absence of unnecessary risks, of a match of a system's operation to the system's context, and the health of the system's information flows.

In the domain of Automated Driving, where part of our research and development is situated, we see these concepts in various ways. Just like in any system, a fault in the power supply of control units or other components will lead to failures of functional units, like camera-based perception, but degradations may already have effects that are more difficult to assess: scratches on a lens, e.g., may lead to lower detection rates of pedestrians in the stream of images that a camera delivers to the AI of the perception system. This lower rate will in turn delay detection, making the system less safe, but not necessarily lead to a direct failure, which would be tragic but at least unequivocal and thus cause for investigations. (See the framework by JAMA (2022) for details on such safety aspects.) The AI, furthermore, can only recognize traffic participants it was trained to classify, so novel transportation modes like unicycles render it outdated – but, again, a failure of detection can only be measured if one of these rare objects is encountered. In these situations, where the system does not meet the AI's requirements, like the lens degradation, or where we face the loss of AI fitness, we face novel challenges to system health management.

First, system health monitoring becomes subject to environmental factors, some of them possibly non-observable or even unknown, which often dominate early warning signals. It is, e.g., possible that the degradation of a camera lens coincides with a seasonal change towards favorable weather conditions, hiding reduced detection probabilities that the perception system would otherwise exhibit. This also illustrates the second challenge: AI-based systems are set to cope with change and to adapt to an open world. Therefore, the overall process that generates observable data for health management is an interwoven combination of system and environment, each with its own dynamics and timelines, and not all of them fully observable. Third, system health monitoring needs to handle uncertainty and probabilistic effects: As Marcus and Davis point out in their critique on current AI approaches (2019), typical algorithms produce only "rough-ready" results and even miniscule differences in the input might lead to misclassifications or other detection failures (Eykholt et al. 2018) without this being an actual system health issue.

Furthermore, embedded AI challenges health management due to the ambivalence in the target of any (preventive) maintenance action or intervention: The system's AI parts might need an update to fit changes in the world or upgrades of the system, but also the non-AI parts might need to be maintained such that they continue to fulfill the AI's requirements. Borth and van Gerwen described this for the comparable dynamic developments that hamper digital twins (2019), stating that "dynamics of a system's environment can render the underlying model unfit w.r.t. the changing reality," leading Pileggi et al. to develop the concepts for lifecycle management that addresses lack of fitness due to change or due to model insufficiencies alternatively (2020). The machine learning and data science communities understand the resulting need to maintain their models, but as part of system lifecycle management, this continues to be a challenge, as, e.g., Almazrouei et al. state in their review of the advancements of artificial intelligence-based models for predictive maintenance (2023).

Joining the probabilistic nature of embedded and cyber-physical AI with its dependency on operational conditions, these challenges led us to develop system health analytics based on probabilistic reasoning that spans environment, AI functionality, physical components, information flow timeliness and quality, as well as system fitness. In this article, we extend our approach such that it considers expectations, i.e., necessary success rates and needed distributions over complex system variables next to direct observations on system states or performance indicators. We describe our resulting system fitness modelling and its basis in probabilistic reasoning in Section II before we demonstrate its diagnostic power in Sections III and IV for performance diagnostics and then for AI assessments. Section V explains how to then use it for prognosis and its application in lifecycle management before we conclude this contribution.

2. SYSTEM FITNESS ASSESSMENT MODELS

Building on the insight that systems with embedded AI are characterized by their information flows and their capability to process relevant information at the right place and time to enable decision-making, we express these aspects and factors impacting them in symbolic variables together with an encoding of their relationships. While such models could use several representations, we chose to use Bayes nets (Pearl 1988) for their capability to execute efficient probabilistic computations. Consequently, our models represent dependencies as edges in a directed graph together with conditional probability distributions over parent–child relationships that follow causal, functional, or probabilistic functions, defining the joint probability distribution over all variables via the chain rule. As causal modeling leads to modularity (Pearl 2009), supporting the use of probabilistic programming (Pfeffer 2016), this approach enables scalable model realization for large and complex systems. Here, we follow systematic generation procedures, as shown, e.g., at Daimler (Borth and Von Hasseln 2002) for the purpose of diagnosis and NASA (Ricks and Mengshoel 2009) for both diagnostics and large-scale criticality and risk assessments. Our own work covers the same applications, with (Gerwen et al. 2024) showing the latest extension towards differential diagnostics with interventions, and (Borth and Barbini 2019) illustrating mission readiness prognosis.

In this article, we adopt the Bayes net shown in Fig. 2 for our health analytics of a camera-based perception system for automated driving that serves as an example. The network follows the conceptual causal structure illustrated in Fig. 1. It covers operational conditions of the automated vehicle (AV), key hardware like Electronic Control or Compute Units (ECUs), forming a self-contained extract of the larger Bayes net used in our productive work, with a selected example of quality of information reasoning. On the latter, the network uses the image contrast in a camera-based perception system as starting point, which is a dominant factor in the object detection's likelihood of success in such systems. The full assessment model covers other sensors as well, but also, for the camera section, additional quality aspects, like sharpness, object occlusion, etc.

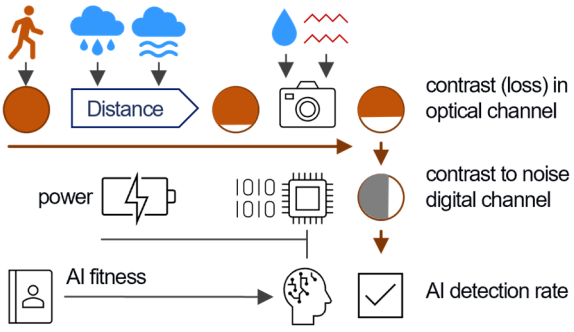


Figure 1. Causal flow in camera-based perception.

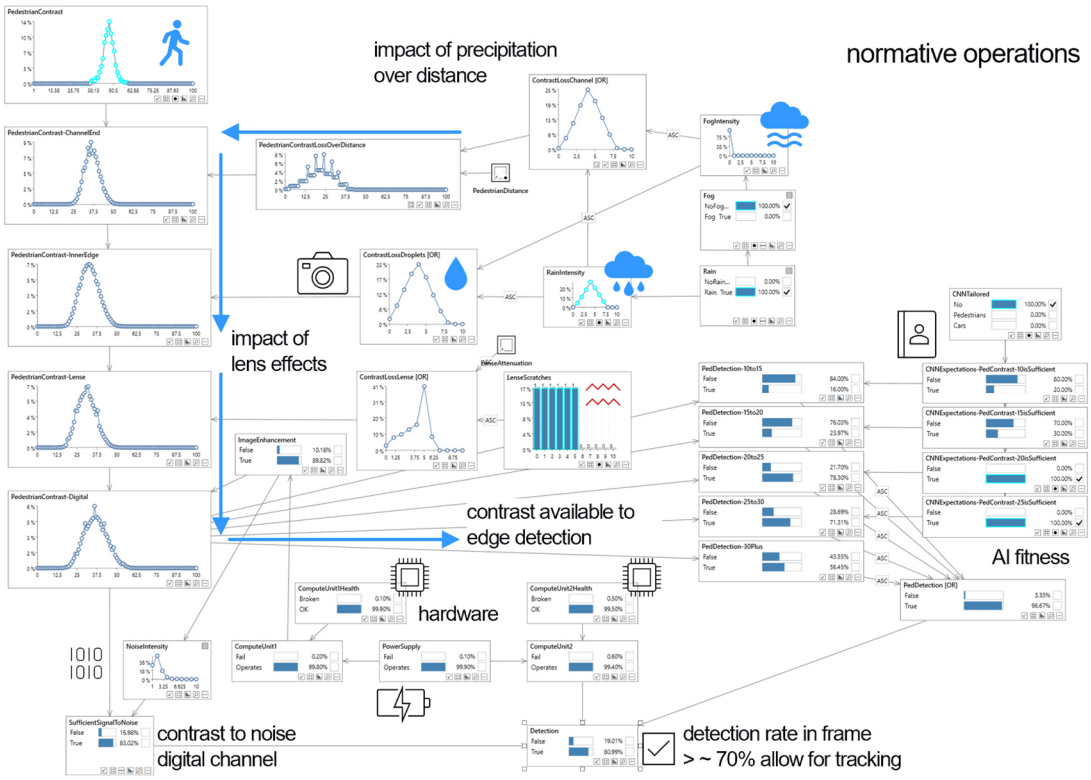


Figure 2. Bayes net assessing fitness of AI-based system. The figure shows an annotated screenshot from Bayes Server.

2.1. Experimental Network Calibration

Next to encoding the physics-informed parts of the Bayes net, we need to establish information about the AI, i.e., about the deployed convolutional neural networks (CNN). This task of network calibration can typically only be done with experiments, as CNNs are sub-symbolic and therefore do not offer understandable rules about their behavior. In our work, we gather experimental data in a controlled simulation environment or from data sets available in the Automated Driving domain. For the former, we use CARLA (Dosovitskiy et al. 2017) for which Fig. 3 shows examples with two weather conditions. For the latter, we work on the Virtual KITTI 2 data set (Gaidon et al. 2016) (Cabon et al. 2020).

Figure 3. Camera view (left) and scenes in CARLA.

3. PERFORMANCE DIAGNOSTICS

The Bayes net shown in Fig. 2 supports prognostic reasoning from causes to effects, e.g., predicting a loss of performance under adverse weather conditions, and diagnostic reasoning from observations to probable causes. Regarding the latter, a well-established usage of such networks is failure diagnosis, where, e.g., a loss of functionality is observed as evidence *e*, with the Bayes net then inferring the likelihood of possible causes *H* given evidence *e* according to the Bayes rule $P(H|e) = [P(e|H) * P(H)] / P(e)$, with $P(H)$ the initial prior belief in *H*, $P(e)$ (evidence) the total probability of e, and $P(e|H)$ (likelihood) the probability of evidence *e* given *H*.

Such a loss of functionality would typically be entered in as hard evidence, i.e., an observation of a specific state, e.g., that it works or doesn't work. However, there is also the option to enter soft evidence. Bayesian inference uses soft evidence to account for unreliable or noisy sensors or expert judgments, expressing uncertain observations, but also for evaluating observed distributions rather than a single fixed state. In our use case, this allows us to observe and to reason about detection rates in the perception system of AVs, a key performance indicator, but also to express how the quality of an information flow manifests, e.g., that the signal-to-noise ratio for the contrast-based edge detection is often above sufficiently high thresholds, but sometimes, with a certain likelihood, below. We regard the resulting diagnostic reasoning as performance diagnosis.

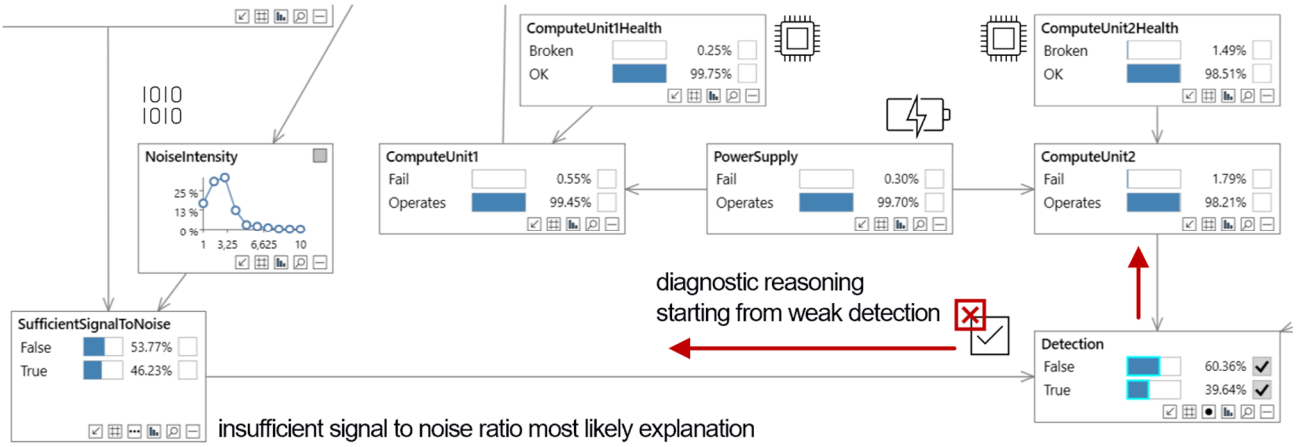


Figure 4. Diagnosis based on failure rate

Performance diagnosis is crucial for AI fitness assessments, as we explain in the next section, but other, more classical use cases exist: The beforementioned work of Gerwen et al. introduces production printing as one of its use cases, where performance diagnosis (which is outside the scope of that publication), can be used to conclude that *a high rate* of paper ejections – a transient fault, expressed in a ratio and thus equivalent to a likelihood – indicates either wrong data about the paper in use or a paper conditioning unit that is unable to deliver pages of the right temperature and levels of moisture to the paper path. This, in turn, might be due to hardware or software faults, or the use of the printer outside its operational domain, e.g., in a building that is too wet – a root cause that is like environmental challenges of AVs.

In that latter domain, we see the suitability of performance diagnosis to determine if a system meets the expectations of its AI, or, more formally, if a system which is used in its operational design domain (ODD), i.e., the operational context and environment for which it was designed, provides information of sufficient quality and timeliness to the AI components such that the system's goals can be met.

Fig. 4 provides an example for this in the cutout of the network above, as a diagnostic process starting with too low detection rates concludes that the pathway does not meet signal-to-noise expectations for >53% of the input flow, indicating a lens or a signal processing issue.

This diagnosis, however, came with some caveats or conditions: We stated that the rain conditions were acceptable, i.e., within the so-called operational design domain (ODD), visible in the node RainIntensity in Fig. 2 that attributes no likelihood to its states for intense rain at the right of its scale. We also ruled out intense scratching on the camera LensScratches, which can be based already on a visual inspection, and, most importantly, stated that the CNN at the core of the AI algorithm in use typically works when the contrast-levels in the input images reach at least 20 on a [0,100] scale (AI fitness nodes).

Cutout from the Bayes net in Fig. 2 with extra evidence.

That last piece of information offered to the diagnosis was essential: given that the present rain would not reduce the contrast of the pedestrian against its background below acceptable levels over the distance between the vehicle and the pedestrian, the issue had to be noise or, rather unlikely, a hardware glitch. It stems from experiments in the setup described in Section 2.1 and depends on the AI algorithm. The AI algorithms we work with are from the YOLO series of real-time object detection systems, first introduced by Redmon et al. (2016). As various versions of YOLO exist in this series, we can use them to realize perception systems with different strengths and weaknesses, thus providing a setting to evaluate AI fitness. (Compare, e.g., Fig. 6 to see how even related CNNs can differ in their performance.)

4. CYBER-PHYSICAL AI FITNESS ASSESSMENTS

Our system health analytics for cyber-physical and embedded AI, Fitness Assessment for short, consists of sets of investigative steps we form from the parts introduced above. The pattern we follow in this is that we trace information flows while checking if contributing parts meet expectations. As the observability of the various flow parts is different, this might result in diagnostic assessments that are direct comparisons between expected and observed distributions, but also in assessments that use extended network fragments that handle parts of the system as a black or grey box. Furthermore, we ask the reader to consider that the steps we follow are based on observability – and any transfer of our work into another setting will require some adjustment if there are differences in this.

Another point to stress is that we cannot and do not assume that the ground truth is known for parts of the reasoning unless (i) we switch from the actual open-world setting to controlled environments, like the software in the loop simulations with CARLA or (ii) we annotate data to determine the ground truth, as it was done for the Virtual KITTI data sets. We detail the consequences of this in the next section.

4.1. Expectations in an Open World

Assessing the health and fitness of AVs translates to a process that monitors operational systems and draws conclusions from their performance in an open world. In this, there is a discrepancy to other diagnostic tasks, as we do not know for certain how the world looks except for the data of the very system that we assess. Therefore, we face the situation that we, e.g., expect the perception system to detect and locate all pedestrians the AV encounters – without knowing how many there are and if any discrepancy between expected and actual performance may be explained away by mitigating circumstances. We cannot, e.g., state that there were 10 people on the road and diagnose perfect system health as the AI found them all. Moreover, while we could state that we expect to encounter 8–10 pedestrians in a road section of a certain type in a given time slot, and therefore would expect to detect as many, seeing only 4 might be due to occlusions as well as simply due to chance and thus not a perception issue.

While this does not change the diagnosis of an actual failure – like overlooking and harming a pedestrian – we therefor need to assess the health of the systems under this lack of ground truth data. One way of mitigating this constraint is the law of large numbers: random effects like the afore-mentioned occlusion vanish in the statistics of the many, allowing us to check our expectations of absolute numbers of detected pedestrians in regular samples: if they drop, the world changed or the system is compromised – and we can differentiate between these possibilities easily with offline processes like fleet analytics or other comparisons.

Faster and within the system – and thus available for live health assessments – is a comparison between different points in time and possibly across sensor systems. If, e.g., the camera-based perception sees an object at 10m distance but fails to consistently detect it at 20m in situations where our expectation was that it should have been seen (with no mitigating factors like heavy precipitation), we conclude that there is an issue, as our expectations were not met. This latter line of reasoning establishes its own ground truth in hindsight: Given that the perception system found a pedestrian, we check if the process leading to its detection indicates a healthy information flow.

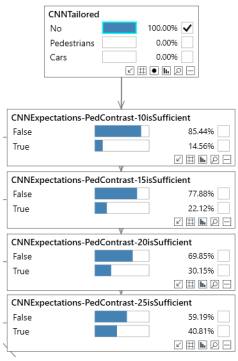
In its most extreme form, this allows us to assess the AI-based system's performance in total, i.e., from input to output without white-boxing the rest of the perception pipeline: Using a short-term memory of key performance indicators like the percentage of frames in which the algorithm's confidence of a detection was sufficiently high and the trajectory data of the AV, we reconstruct the detection rates *given various distances* between pedestrian and AV, and judge the system's health based on the assessment of the Bayesian network, i.e., the expectations visible via its reasoning.

4.2. Diagnostic Assessments

With the notion of expectation on both the open world's states and the performance of AI-based systems in it, we use the probabilistic inference offered by the Bayesian network in diagnostic fitness assessments of cyber-physical AI.

Starting with the understanding that the performance of AI relies in part on the quality of the information made available to it, we have a strong entry point for diagnosis, as that quality is observable. In our example, we instantiate the Contrast Digital and NoiseIntensity nodes with observed distributions and set nodes capturing the environment like Rain/FogIntensity to values representing conditions during the observed time. As the network propagates the evidence about the information quality both upward and downward, it diagnoses the lens effects in the upward pathway, conducting a performance diagnosis, and likely concluding that lens scratches scattered the light if low contrast values were not explained away by precipitation or diagnosing a good lens if those values met expectations. Next, the network assesses the status of image digitalization and enhancement, which is foremost a direct check between incoming contrast and observable noise. If the observed noise levels are according to expectations, the network infers that the involved software and hardware functions and infers the opposite along the lines of Section 3 if not.

Overall, this creates two possible scenarios if the detection rate observed in Detection was too low to start with: if the observed rate is consistent with observed low information quality, the diagnostic reasoning described above pinpoints causes for that, and we state that AI is likely functional, but the perception system does not meet the operational quality to meet the AI's expectations. If, on the other hand, there is no sufficient explanation for the lack of performance in the observed information flows, we state that the AI core itself does not meet expectations: it either requires higher quality information as it should – or its training data does not adequately represent the objects in the world, e.g., if there were no scenes with children presented to the algorithm. Fig. 5 shows an example of the former. If it can be ruled out, e.g., with dedicated tests, the latter diagnosis remains.



Diagnostic assessment of which contrast levels are sufficient for pedestrian detection. The Bayes net infers that even a level of 25 from 100 does not suffice for detection with a ~60% likelihood.

By comparison, the expectation expressed in Fig. 2 states that 25/100 should always suffice.

Figure 5. AI Fitness Issue. Cutout based on new evidence.

5. PROGNOSIS

The full capabilities of our approach for the lifecycle health management of AI-based systems come together with the Bayesian network's *predictive* reasoning powers. Here, we distinguish between two approaches: assuming future states within the causal scope of the current fitness assessment and counterfactual reasoning, which captures interventions including upgrades or updates to the hardware and software of the system. In combination, these techniques predict both expected developments and arbitrary what-if scenarios, allowing the prediction of a lack of fitness and safety before it becomes likely, thus enabling us to take actions in time.

5.1. Predicting Foreseeable Health and Fitness States

Many of the causal factors that our fitness assessment model considers change over time within the scope of a predefined state space that was set during the generation of the model. The form and functions underlying this change vary greatly. Weather patterns, e.g., change their a priori likelihoods over the seasons: heavy weather is always possible, just more likely in fall. Physical degradation D, on the other hand, will (without intervention) never get better and follow a function of time t like, e.g., $D(t) = D_0 + k \times t^n$, with D_0 the initial state of degradation, k a degradation rate constant, n an empirical exponent, e.g., $n = 1$ for linear processes.

We then predict future health and fitness states by starting with a diagnosis that estimates the current states of relevant components, like lens degradation in our example. With this, we gain estimates of the respective D_0 values, allowing us to predict their future degradation $D(t)$, by formula, if known, based on data-driven remaining useful life estimates or also within Bayesian models, offering several advantages for this task, as laid out by Hostens et. al (2024). Entering distributions for these $D(t)$ values into the assessment Bayes net as soft evidence, thus expressing the uncertainty that we face concerning their precise future value, we gain a model of the future state of the AI-based system. Next, we compute the risks that arise from operating the system in this state: the expected overall performance – and thus also the likelihood of failures – follows from adjusting the a priori likelihoods of the remaining factors towards the future at time t, e.g., by tuning precipitation expectations to the upcoming season. We can also make the risk assessment explicit by setting the expected degradation and asking the Bayes net under which circumstances this degraded system will fail and calculating the likelihood of those conditions happening. If the resulting forecast includes non-acceptable risks, we take actions, like switching out the camera, knowing that this is necessary while avoiding costs if not.

In our opinion, this line health management is mandatory in domains where AI solves safety-critical applications, like Automated Driving, as it encompasses risk assessments, thus surpassing approaches that are purely based on the remaining useful life of components.

5.2. Prognosis of What-if Scenarios

While powerful, the prediction of system health and fitness described in Section 5.1, is bound to the causality which is encoded in the Bayesian model. However, there are ample scenarios within lifecycle management that can change this causality. Hardware replacements like that of an AV's camera against an upgraded model may, e.g., reduce the noise introduced into the signal pathways, changing a highly relevant impact factor. Similarly, an update to the AI can change detection likelihoods given certain objects or under certain conditions, like precipitation. Analyzing such 'what-if' scenarios is the domain of counterfactual reasoning, with (Pearl 2009) providing the leading computation framework, Weisberg and Gopnik (2013) explaining the value of this reasoning method in our thinking, and, e.g., Andringa, Baptista and Santos (2025) detailing its use in remaining useful life analyses.

Counterfactuals are different from what we described before as they change the Bayesian model to investigate an alternative causality – either via Pearl's *do(x)* operator or by replacing parts of the model to reflect the alternative. In our work, we are most interested in using them to identify and validate necessary changes, e.g., to investigate whether a new version of an AI algorithm may be expected to perform at least according to our expectations. Fig. 6 shows an example of the impact of such a switch between algorithms, depicting significant better detection likelihoods using a larger (and slower) version of YOLO within a controlled virtual environment. Using data like this to determine alternate conditional probability distributions in the AI parts of our assessment model, we create the basis for 'what-if' assessments that analyze if an AI update would indeed lead to a system that meets our expectations, guiding lifecycle management of the AI part, but allowing also the identification of areas of doubtful performance that helps in verification and validation (Paardekooper et al. 2024).

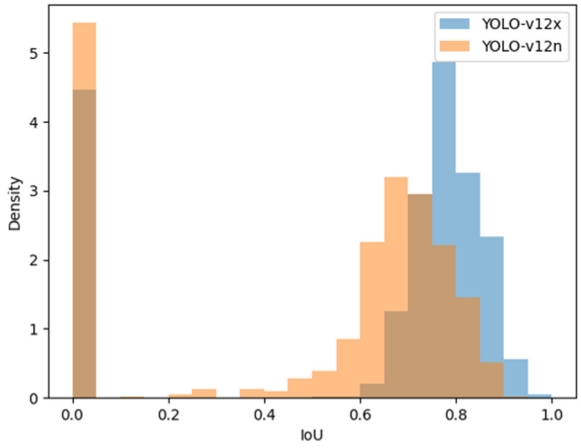


Figure 6. Detection performance comparison measured in Intersection over Union (IoU, high is better) between YOLO algorithms for pedestrians at 19.5-20.5m distance, based on simulations at 223 locations in our CARLA environment.

6. CONCLUSION

Asserting that a system meets expectations seems to be a weak statement in comparison to establishing absolutes like that it is not broken, that there is no wear and tear, that we will see no failures. Yet, for the domain of cyber-physical AI and embedded AI, realizing adaptive behaviors, it is one way for lifecycle management to ensure continuous performance and safety of autonomous and automated systems. In this, it is a necessary way forward – most likely not the only one, but one that addresses the issue that such advanced systems break away from many traits practitioners previously based system health diagnosis and prognosis methods on, like repeatability, controllability, or testability. We achieve this chiefly by combining two lines of thought:

First, expectation-based system health analytics incorporate that the performance of systems with strong AI inside is highly context dependent as well as subject to chance, i.e., effects and processes that are for all practical effects partly probabilistic. It takes on the fact that a system's health and fitness show only in success rates and no longer in absolutes and forms its expectations in this fashion: a healthy system is one that is, for the foreseeable future, fit to deliver acceptable rates of success under the circumstances that one might reasonably expect.

Secondly, our system health analysis and prognosis for embedded and cyber-physical AI establishes expectations in all relevant relationships: from the AI towards the quality and timeliness of the system's information processing given the state of the environment, from the system's software functions towards the AI, but also from the AI-based system towards the happenstances in the operational domain.

Given our implementation choice to assess the fulfilment of the breakdown of all these expectations with Bayesian reasoning, this allows us to provide for diagnosis as well as prognosis – thus providing, as stated in the opening of this section, a way to ensure continuous performance and safety, which is essential for the trust in systems that we empower to make decisions.

REFERENCES

Andringa, J., Baptista, M. L. & Santos, B. F. (2025). Counterfactual Explanations for Remaining Useful Life Estimation within a Bayesian Framework. *Information Fusion* 118. doi.org/10.1016/j.inffus.2025.102972

Japan Automobile Manufacturers Association (JAMA) (2022). *Automated Driving Safety Evaluation Framework*. www.jama.or.jp/english/reports/framework.html

Borth, M., & Barbini L. (2019). Probabilistic Health and Mission Readiness Assessment at System-Level. *Proceedings of the Annual Conference of the PHM Society 11*(1). doi.org/10.36001/phmconf.2019.v11i1.777

Borth, M., De Oliveira Filho, J., & van der Ploeg, C. (2024). Fitness Assessment of AI-Based Systems. *Prognostics and System Health Management Conference (PHM)*, 235–40. doi.org/10.1109/PHM61473.2024.00050

Borth, M., & van Gerwen, E. (2019). Tracking Dynamics in Concurrent Digital Twins. In *Complex Systems Design & Management*, edited by Bonjour, Krob, Palladino, & Stephan. Springer International Publishing. doi.org/10.1007/978-3-030-04209-7_6

Borth, M., & von Hasseln, H. (2002). Systematic Generation of Bayesian Networks from Systems Specifications. In *Intelligent Information Processing*, edited by Musen, Neumann, & Studer, vol. 93. IFIP Advances in Information and Communication Technology. Springer. doi.org/10.1007/978-0-387-35602-0_14.

Cabon, Y., Murray, N., & Humenberger, M. (2020). Virtual KITTI 2. arXiv:2001.10773. arXiv, January 29. doi.org/10.48550/arXiv.2001.10773

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An Open Urban Driving Simulator. *Proceedings of the 1st Annual Conference on Robot Learning*, 1–16. https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Eykholt, K., Evtimov, I., Fernandes, E., et al. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. (*2018) IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1625-34. doi.org/10.1109/CVPR.2018.00175

Gaidon, A., Wang, Q., Cabon, Y., & Vig, E. (2016). VirtualWorlds as Proxy for Multi-Object Tracking Analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4340-49. doi.org/10.1109/CVPR.2016.470

van Gerwen, E., Barbini, L., Borth, M., & Passmann, R. (2024). Efficient Differential Diagnosis Using Cost-Aware Active Testing. *International Journal of Prognostics and Health Management* 15 (3). doi.org/10.36001/ijphm.2024.v15i3.3849

Hostens, E., Eryilmaz, K., Vangilbergen, M., & Ooijevaar, T. (2024). Bayesian Networks for Remaining Useful Life Prediction. *PHM Society European Conference* 8 (1): 1. doi.org/10.36001/phme.2024.v8i1.4019

Marcus, G. F., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. First edition. Pantheon Books.

Almazrouei, S. M., Dweiri, F., Aydin, R., & Alnaqbi, A. (2023). A Review on the Advancements and Challenges of Artificial Intelligence Based Models for Predictive Maintenance of Water Injection Pumps in the Oil and Gas Industry. *SN Applied Sciences* 5 (12): 391. doi.org/10.1007/s42452-023-05618-y

Paardekooper, J-P., & Borth, M., (2024). Toward a Methodology for the Verification and Validation of AI-Based Systems. *SAE International Journal of Connected and Automated Vehicles* 8 (1). doi.org/10.4271/12-08-01-0006

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press. doi.org/10.1017/CBO9780511803161

Pfeffer, A. (2016). *Practical Probabilistic Programming*. Simon and Schuster.

Pileggi, P., Lazovik, E., Broekhuijsen, J., Borth, M., & Verriet, J. (2020). Lifecycle Governance for Effective Digital Twins: A Joint Systems Engineering and IT Perspective. *IEEE International Systems Conference (SysCon)*. doi: 10.1109/SysCon47679.2020.9275662

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640. arXiv, May 9. doi.org/10.48550/arXiv.1506.02640

Ricks, B. W., & Mengshoel, O. J. (2009). Methods for Probabilistic Fault Diagnosis: An Electrical Power System Case Study. *Annual Conference of the PHM Society* 1 (1). papers.phmsociety.org/index.php/phmconf/article/view/1594

Weisberg, D. S., & Gopnik, A. (2013). Pretense, Counterfactuals, and Bayesian Causal Models: Why What Is Not Real Really Matters. *Cognitive Science* 37 (7): pp. 1368-81. doi.org/10.1111/cogs.12069