Rafael Machado Peluzio

Análise dos modos de falhas em sistemas de Veículos Autônomos

Viçosa, MG

2022

Rafael Machado Peluzio

Análise dos modos de falhas em sistemas de Veículos Autônomos

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 402 – Projeto de Engenharia II – e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador Prof. Heverton Augusto Pereira

Viçosa, MG

2022

Rafael Machado Peluzio

Análise dos modos de falhas em sistemas de Veículos Autônomos

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 402 – Projeto de Engenharia II – e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Trabalho aprovado em 12 de dezembro de 2022.

COMISSÃO EXAMINADORA

Prof. Heverton Augusto Pereira
Orientador

Prof. Rodolpho Vilela Alves Neves
Membro Avaliador

Prof. Victor Pellanda Dardengo
Membro Avaliador

Viçosa, MG
2022

Agradecimentos

Este trabalho representa para mim não apenas o meu trabalho de conclusão de curso, mas também a conclusão da realização de dois grandes sonhos meus. Em primeiro lugar, o sonho de tornar-me engenheiro eletricista, e em segundo lugar, o sonho de realizar um intercâmbio e vivenciar outras culturas.

Agradeço ao professor Heverton Pereira, pela orientação na elaboração deste trabalho, e ao professor Nicolae Brinzei, quem orientou este estudo como parte de meu PFE.

Sou grato ao acompanhamento das pesquisadoras do CEA List, Morayo de Diana, cuja missão me inspirou durante o desenvolvimento deste estudo.

Pela realização destes sonhos, tenho a agradecer a todas as pessoas maravilhosas que conheci ao longo desta jornada. Todos os colegas de curso que me apoiaram até este momento, todos os amigos que fiz em Viçosa e todos os momentos que passamos juntos e que me permitiram sobreviver às provações deste período. Menciono aqui especialmente Marcella, Iure e Karina.

Aos amigos do Cool do Charmois, Ana Clara, Marco Antônio, Felipe e Lucas que fizeram minha morada na França um ambiente amigável e tranquilo.

Aos amigos do Lacratop, Heitor, Arthur Henrique e Rodrigo Eduardo.

E em último e mais importante lugar, a minha família, que sempre foi a minha forte fundação. Meu pai João Batista, minha mãe Telma e meu irmão Lucas.

Resumo

Este trabalho faz uma revisão das normas vigentes à confiabilidade da nova geração de veículos, em relação à aplicação de tecnologias que integram Inteligência Artificial. São apresentados os métodos clássicos de análise de falhas, as falhas específicas aos veículos autônomos e a limitação da verificação destas pelos métodos clássicos. Novos métodos de correção e prevenção destas falhas são apresentados, como STPA, PEGASUS e a aplicação de visão computacional na análise de cenários.

Keywords: FMEA, HAZOP, SOTIF, STPA, Inteligência Artificial, Veículos Autônomos.

Abstract

This work presents a review of the current standards applied for the reliability of the new vehicle generation, related to the application of technologies that integrate Artificial Intelligence. Classic fault-analysis methods are presented, aswell as specific failures of autonomous vehicles and the limitations of the verification of such failures by the classic methods. New methods for the analysis of these failures are presented, such as STPA, PEGASUS, and scenario-based approaches.

Keywords: FMEA, HAZOP, SOTIF, STPA, Artificial Intelligence, Autonomous Vehicles.

Lista de figuras

List of abbreviations and acronyms

ABNT	Associação Brasileira de Normas Técnicas
SAE	*Society of Autonomous Engineers*(Sociedade de Engenheiros Automotivos)
FTA	*Fault Tree Analysis*(Análise de árvore de falhas)
FMEA	*Failure Modes and Effects Analysis*(Análise de modos de falhas e efeitos)
HAZOP	*Hazard and Operability analysis*(Análise de perigos e operabilidade)
STPA	*System's Theoretic Process Analysis*(Análise do processo teórico do sistema)
SOTIF	*Safety of the Intended Functionality*(Segurança da funcionalidade pretendida)
IA	Inteligência Artificial
AI	*Artificial Intelligence*(Inteligência artificial)
DNN	*Deep Neural Networks*(Redes Neurais Profundas)
PEGASUS	*(Project for the Establishment of Generally Accepted quality criteria, tools and methods as well as Scenarios and Situations*(Projeto para o Estabelecimento de Critérios de Qualidade, ferramentas e métodos, além de cenários e situações geralmente aceitas(Para a confiabilidade da condução autamente automatizada))
ISO	*International Organization for Standardization*(Organização internacional para a padronização)
HARA	*Hazard Analysis and Risk Assessment*(Análise de perigos e avaliação de riscos)
SOTIF	*Safety of the Intended Functionality*(Segurança da funcionalidade pretendida)

Sumário

1 Fundamentação Teórica

1.1 Níveis SAE

Os níveis SAE são os níveis de automação de condução estabelecidos pela SAE, órgão responsável pelas regulamentações internacionais para automobilismo de consumo. A tabela explicativa dos 5 níveis da SAE é apresentada na figura 1, seguida da descrição de cada nível e da análise do atual nível de automação na indústria(SAE, 2014)(RAJABLI et al., 2020).

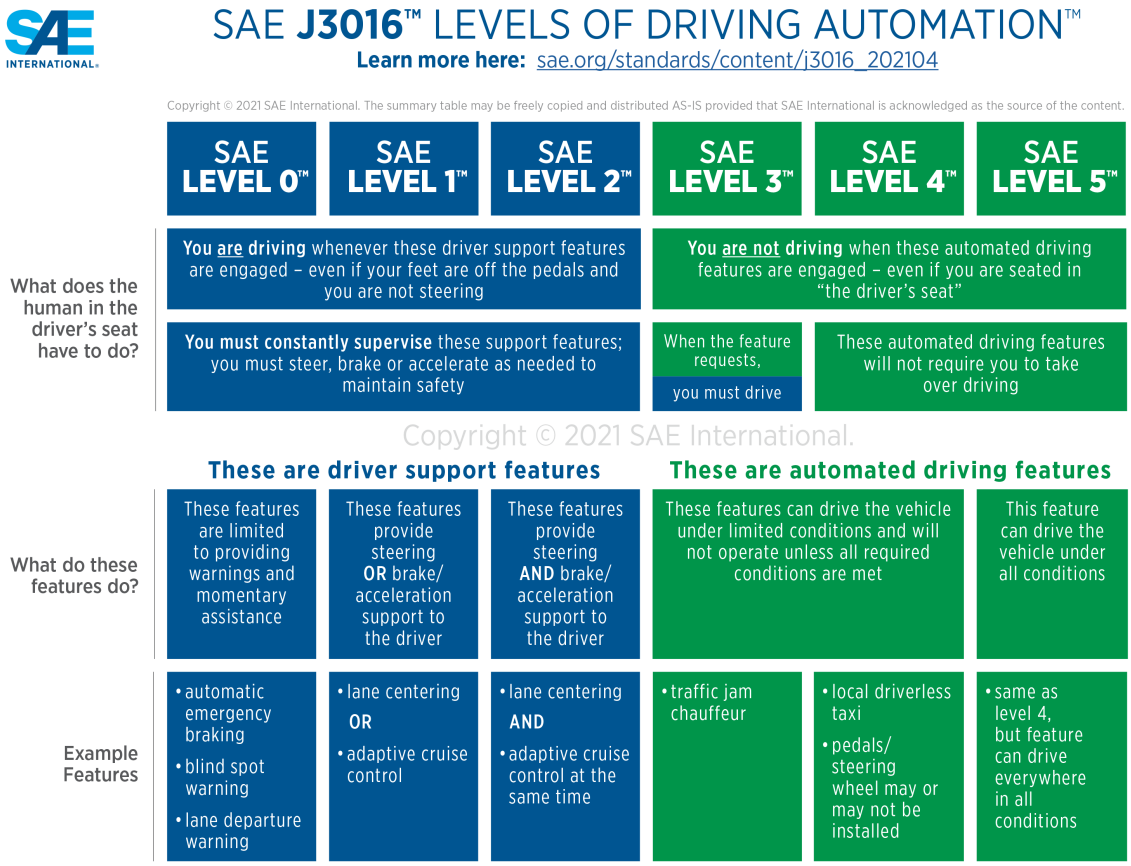

Figura 1 – Tabela de níveis SAE de automação da condução (SAE, 2014)

Esses níveis classificam a autonomia do veículo em uma escala de 0 a 5. O nível 0 representa um veículo cujas funcionalidades automatizadas são extremamente limitadas, havendo apenas alguns indicadores para os sistemas de assistência ao condutor e auxílio ao controle da condução. Por exemplo, funções básicas como ABS, aviso de ponto cego ou aviso de saída de faixa.

Os níveis 1 e 2 representam um veículo capaz de intervir na aceleração e/ou travagem em situações específicas do funcionamento normal do veículo, notavelmente para facilitar a tarefa do condutor. Podemos citar a assistência de permanência em faixa e o controle de cruzeiro adaptativo.

O nível 3 é onde a condução automatizada já é uma característica principal do veículo. A condução é feita automaticamente, mas quando a situação o exige, o condutor deve intervir na condução. Este é o nível atual de veículos de condução autônoma como Tesla.

O nível 4 é o nível em que não é necessário usar um motorista para garantir uma condução segura e onde o veículo dirige absolutamente sozinho. A única distinção entre os níveis 4 e 5 é que no nível 4 o veículo pode ter certas restrições sobre onde pode ou não ir. Por exemplo, o veículo pode ser limitado a dirigir na rodovia ou em cidades que possuem um nível de infraestrutura necessário, enquanto o nível 5 afirma que onde quer que um humano possa ir com um veículo, o veículo também pode ir sozinho.

1.2 A utilização de Inteligência Artificial em Sistemas Autônomos

A inteligência artificial, ou IA, tem várias definições. Na década de 1950, o teste de Turing era o padrão-ouro para determinar a inteligência de uma máquina. Este é um teste durante o qual uma máquina teve que ficar irreconhecível durante uma discussão com um humano. Em termos gerais, a IA pode ser definida como a inteligência demonstrada por máquinas. A IA imita as funções cognitivas humanas e as técnicas de resolução de problemas, obtendo na maioria dos casos um desempenho melhor(BRABAND et al., 2020).

Os algoritmos de IA conseguem automatizar muitas tarefas que exigem inteligência, como processamento de dados e reconhecimento de imagens, amplamente utilizados em veículos autônomos. Contudo, para descrever as tecnologias de IA utilizadas em veículos autônomos, faz-se necessário fazer algumas distinções relativas à IA:

1.2.1 Aprendizado vs Conhecimento

IA baseada em conhecimento é criada tentando reproduzir o comportamento de um especialista em um determinado domínio do *software*. É uma estratégia adequada para obter resultados rápidos quando a tomada de decisão do especialista pode ser descrita de forma suficientemente simples.

O sistema de travagem anti-bloqueio (ABS) é um bom exemplo de IA baseada no conhecimento com um enorme impacto atualmente. O ABS Baseia-se em técnicas anteriormente utilizadas por motoristas profissionais e as implementa automaticamente.

No entanto, as IAs baseadas no conhecimento apresentam alguns problemas para a condução autônoma, principalmente devido à dificuldade de automatizar certas tarefas como o reconhecimento de imagens. Além disso, um dos objetivos dos veículos autônomos é superar a tomada de decisão humana e aumentar o nível de segurança desses sistemas.

A IA baseada em aprendizado, por outro lado, é gerada pelo treinamento de um algoritmo usando cenários didáticos, ou bancos de dados, para executar uma determinada função. A maior vantagem dos algoritmos de aprendizado de máquina é sua capacidade de melhorar seu funcionamento com dados coletados de sistemas operacionais reais. No entanto, sua complexidade é maior em comparação com sistemas baseados em conhecimento, gerando uma dificuldade de compreensão.

Essa dificuldade de compreensão também pode ser descrita usando uma comparação entre um sistema de IA e uma caixa preta. É possível que os engenheiros entendam quais dados foram alimentados ao algoritmo e quais dados foram extraídos dele, mas os cálculos específicos, aproximações e métodos que estes usam permanecem geralmente ocultos, e quanto maior a complexidade, mais é difícil de entender os algoritmos.

A IA baseada em aprendizado já demonstrou capacidades superiores aos humanos em muitas atividades, como nos jogos de tabuleiro, xadrez e GO, ou até diagnóstico de câncer, como fez a IA da IBM, Watson(ZARKADAKIS, 2019).

1.2.2 Aprendizagem Supervisionada, Não Supervisionada e por Reforço

A IA baseada em aprendizado lida com problemas de classificação e regressão, ou seja, pega os dados e os classifica com base em outro conjunto de dados, ou prevê o comportamento com base nas informações disponíveis. No entanto, essas três abordagens diferem em particular porque o aprendizado supervisionado, como o próprio nome sugere, permite que o algoritmo saiba as respostas corretas a dar durante a análise dos dados, o que significa que ele aprenderá rápida e exatamente como replicar o comportamento que a ele foi mostrado, mas provavelmente não desenvolverá características adicionais, como abstração de dados mais profunda.

O aprendizado não supervisionado, por outro lado, é usado quando o algoritmo de aprendizado não recebe uma resposta positiva ou negativa às suas avaliações. Este tipo de algoritmo é usado para encontrar significado nos próprios dados originais.

O aprendizado por reforço é uma técnica na qual o algoritmo recebe feedback positivo ou negativo de forma iterativa, dependendo de seu desempenho, mas mantendo a liberdade de criar soluções para o problema dentro de determinadas restrições.

1.2.3 Aplicações da IA a sistemas Não-Críticos

Os sistemas não críticos, ou seja, sistemas cujas falhas ou erros não são susceptíveis de causar perigo imediato a pessoas e bens, possuem produtos que aplicam IA no mercado há algum tempo. Até muito recentemente, eram quase sempre na forma de produtos e serviços indiretos, como algoritmos de busca, mas recentemente ganharam uma dimensão mais física, integrando-se, por exemplo, em sistemas como assistentes pessoais, drones de entrega e robôs de limpeza.

Nestas aplicações, numerosas combinações diferentes dos tipos de IA mencionados anteriormente já foram aplicadas e mostraram resultados concretos.

1.2.4 Aplicação de IA a sistemas críticos

Até o momento, os sistemas de IA não possuem uma certificação de segurança específica que permita sua utilização para equipamentos críticos de segurança. Especialmente no campo automotivo Baseando-se em (BRABAND et al., 2020) (WANG; CHUNG, 2022).

A IEC 61508 é a norma internacional para segurança funcional de sistemas E/E/PE(Elétricos, Eletrônicos e Eletrônico-Programáveis), e é a base utilizada por especialistas para desenvolver a norma ISO 26262(ISO, 2011). esta norma não recomenda o uso de tecnologias baseadas em IA para sistemas relacionados à segurança. No entanto, uma evolução que representa o avanço da indústria rumo à inserção destas tecnologias é o padrão UL 4600(ANSI/UL, 2020), que em 2019 tentou fornecer algum nível de classificação de segurança para estes sistemas. Esta norma forneceu um argumento para uma abordagem de caso de segurança para veículos autônomos, mas não mostrou como realizá-lo com sucesso ou verificar o sucesso de sua implementação.

O padrão UL 4600 inclusive afirma que ”a conformidade com esta norma não é garantia de segurança de veículos automatizados”.

Mesmo que a automação completa da função de condução, ou seja, o alcance dos níveis 3 e 4 da norma SAE, ainda não seja uma possibilidade em termos de segurança, já existem muitas tecnologias críticas para a segurança, principalmente na área automotiva, que utilizam IA.

Alguns desses aplicativos já são usados em sistemas avançados de assistência ao motorista (ADAS):

- *Anti-locking Brake System* - ABS;
- *Automatic Parking Systems* - APS;
- Sinalização das vias e reconhecimento de pedestres;

- Assistência à frenagem;
- Assistência ao mantimento em via;

Esses aplicativos usam métodos e algoritmos diferentes. ABS, assistência à frenagem e as iterações mais básicas do APS são classificados como sistemas especializados porque aplicam algoritmos que imitam determinado comportamento especializado baseado em conhecimento(AULINAS; SJAFRIE, 2021).

No entanto, para poder desenvolver funções mais avançadas, é necessário dotar os sensores do automóvel de capacidades de detecção cada vez mais importantes, capacidades como a transformação da imagem de uma câmera ou de um conjunto de câmeras em informação útil, como a posição de obstáculos ao redor ou a identificação de pedestres, a presença de sinais de trânsito e outros objetos de trânsito em geral(BRABAND et al., 2020).

Para alcançar um alto nível de reconhecimento de padrões, como o necessário para reconhecimento de imagem, redes neurais artificiais foram implementadas para funções ADAS (sistema avançado de assistência ao motorista), como pode ser visto no sistema de assistência à condução da BMW. Esses algoritmos dependem de bancos de dados para serem treinados para realizar determinada tarefa.

No campo de veículos autônomos (AV), as principais tecnologias de IA utilizadas são o *Deep Learning*, que é um tipo de inteligência artificial baseada no uso de várias camadas neurais de mesmo tamanho para realizar operações sequenciais. Os AVs geralmente implementam funções de reconhecimento de padrões, a fim de perceber seu ambiente de maneira física e conceitual e tomar decisões.

Para sistemas de IA, no entanto, é muito difícil verificar se essas funções atendem às especificações de segurança propostas por lei. O aprendizado de máquina requer treinamento para poder reconhecer padrões e esse treinamento é feito utilizando bancos de dados, ou seja, se quiséssemos treinar um algoritmo de aprendizado de máquina para reconhecer pedestres, precisaríamos treinar esse sistema com milhares, senão milhões, de imagens de pedestres para obter a configuração correta. No entanto, o banco de dados pode conter certos vieses, como a presença de um pedestre sendo obscurecida por condições climáticas extremas, ou talvez até a ausência de imagens de pessoas com deficiência, o que tornaria o sistema do carro incapaz de identificar corretamente essas pessoas.

No caso dos AVs e da IA usada em seu design, o ambiente está em constante mudança, pois é impossível modelar completamente todas as situações que o AV possa encontrar em seu caminho, e assim o algoritmo é modificado com base nos dados percebidos pelo produto após o seu lançamento. Um exemplo muito marcante desse fenômeno é o de uma das atualizações lançadas pela Tesla, que mudou completamente o tempo de resposta

do sistema de frenagem do Model 3 com uma simples atualização de *software* (O'KANE, 2018), deixando os clientes em uma situação estranha em que seus veículos mudaram de comportamento de um dia para o outro.

2 Introdução

Tecnologias baseadas em inteligência artificial, a exemplo de assistentes digitais pessoais como Alexa e Siri, mecanismos de busca na web e carros autônomos, estão cada vez mais integrados à vida cotidiana, e embora cheguem ao mercado com inúmeros recursos úteis e valiosos, na discussão sobre produtos com grandes implicações de segurança, consumidores e governos se preocupam se essas novas tecnologias são seguras e funcionam confiavelmente dentro das funcionalidades pretendidas, especialmente quando a IA é usada para controlar sistemas propensos a acidentes graves como veículos autônomos ou máquinas industriais.

Na indústria automotiva, a maioria das fabricantes já integrou tecnologias baseadas em IA em seus melhores produtos, que alcançam os níveis de segurança exigidos pelas normas e leis que regem atualmente o campo. Contudo, esta adequação pode não ser suficiente para garantir a segurança destes sistemas com o uso de tecnologias cada vez mais complexas (RAJABLI et al., 2020).

Um exemplo dessa constante evolução de tecnologias de segurança e das normas no Brasil é a Resolução Nº 717/2017 do CONTRAN, que estabeleceu cronograma de estudos técnicos e regulamentação para 38 itens de segurança, dentre os quais destacam-se os sistemas de Frenagem Automático de Emergência (AEBS - *Automatic Emergency Brake System*) e de Aviso de afastamento de faixa de rodagem (LDWS - *Lane Departure Warning Systme*).

2.1 O problema

Em anos recentes, fabricantes de automóveis como Tesla, e até empresas de tecnologia como Apple e Huawei, têm desenvolvido intensamente a instrumentação e controle autônomo de veículos através do uso de IA, visando alcançar o nível de tecnologia necessário para a automação de veículos de Nível SAE 4 (Society of Automotive Engineers).

No entanto, a IA apresenta uma série de problemas de segurança decorrentes de sua natureza complexa, como a possibilidade que eventos inesperados ocasionem uma falha em seu comportamento, que tornam os métodos tradicionais de avaliação de segurança incompletos. Esses métodos serão explorados com mais detalhes neste trabalho e as abordagens usadas hoje para desenvolver sistemas baseados em IA mais seguros serão mencionadas.

2.2 Objetivos deste trabalho

Este trabalho é produzido na intenção de validar e direcionar um trabalho em desenvolvimento pela empresa público-privada de pesquisa, CEA List, de Nancy/FR, para criar uma plataforma de simulação de falhas e quantificação da segurança de veículos autônomos.

O objetivo principal deste trabalho é apresentar à equipe do CEA List uma proposta original para o software a ser desenvolvido que atenda a alguma necessidade prática da indústria de veículos autônomos.

Na intenção de contextualizar e direcionar este objetivo principal, uma revisão da confiabilidade de sistemas de veículos autônomos que integram inteligência artificial também é realizada, apresentando as normas relevantes e evoluções recentes dos padrões e métodos utilizados pelo setor automobilístico para validar a segurança de seus novos produtos.

3 Métodos de análise da segurança de sistemas

Nesta seção, os principais métodos existentes de análise de segurança e avaliação de risco serão apresentados, descritos e comentados.

3.1 FTA - Análise da árvore de falhas

A análise da árvore de falhas, do inglês *Fault Tree Analysis*, é um método de análise de falhas dedutivo, dito de cima para baixo, no qual eventos perigosos, como uma colisão, são analisados por meio de relações booleanas entre os componentes do sistema, criando uma árvore de falhas, eventos causais que podem então ser reforçados para atingir o nível de segurança desejado.

Graças à sua natureza combinatória, as árvores de falhas podem ser utilizadas como fonte de novas falhas sistemáticas a partir da análise de combinações de eventos.

Esta análise é geralmente realizada em conjunto com um estudo do tipo FMEA.

3.2 FMEA - Análise dos Efeitos e Modos de falhas

Esta análise, do inglês *Failure Modes and Effects Analysis*, é uma análise sistemática em que cada parte é estudada sob a perspectiva de possíveis falhas que poderiam ser causadas ao resto do sistema a partir de uma falha de um de seus componentes. Geralmente é aplicada como forma de obter as associações entre as diferentes falhas do sistema, e para obter a relação entre as falhas locais e globais.

3.3 HAZOP - Estudo de Perigo e Operabilidade

O Estudo de Perigo e Operabilidade, *Hazard and Operability Study*, é um método de análise de risco constituído de uma equipe de especialistas que, a partir de uma lista comum de parâmetros e vocabulário estabelecido por um líder, tentará imaginar diferentes maneiras pelas quais o sistema pode quebrar e/ou exigir melhorias. Cada reunião em um conjunto de parâmetros é chamada de nó e a análise HAZOP é concluída quando um número suficiente de análises/nós é concluído.

3.4 Análise de Markov e Análise Melhorada de Markov

Do inglês *Markov Analysis and Improved Markov Analysis*, esses dois métodos partem do mesmo princípio matemático. Assim como o método FTA, eles são de natureza quantitativa e baseiam-se na descrição do modelo em termos de estados e transições, que, no caso da análise de Markov aprimorada, também pode ser descrita por uma análise de incerteza, usando o método de Monte-Carlo.

3.5 STPA - Análise do Processo Teórico do Sistema

System's Theoretic Process Analysis é um método de análise de segurança de sistemas com foco em interações. STPA é um método desenvolvido por Takuto Ishimatsu e Nancy Levenson em 2010 com a intenção de ser utilizável para encontrar mais defeitos do que a análise FTA e também para ser utilizado antes, durante e após a execução de um projeto (ISHIMATSU et al., 2010).

O STPA demonstrou sua eficácia na detecção de defeitos de um veículo autônomo compatível com ISO 26262(ISO, 2011) em (MAHAJAN et al., 2017), (ABDULKHALEQ et al., 2017).

O STPA é uma abordagem de cima para baixo, ou seja, considera o comportamento dos componentes do sistema para determinar como suas interações podem causar o(s) perigo(s) especificado(s).

3.5.1 Aplicações do método STPA

Baseando-se na aplicação descrita em (MAHAJAN et al., 2017), STPA repousa sobre três conceitos principais:

- As restrições de segurança;
- A estrutura de controle hierárquico;
- Os modelos do processo;

A primeira etapa é a definição do evento perigoso; A segunda etapa é identificar os estados perigosos causados pela falha de controle nas seguintes situações:

- Ação de controle fornecida e não requisitada;
- Ação de controle requisitada e não fornecida;
- Ação de controle fornecida após intervalo inapropriado;

- Ação de controle fornecida por duração inapropriada;

Depois de analisar os requisitos e restrições de cada sistema, a terceira etapa é determinar como cada estado perigoso pode ocorrer e a quarta etapa é desenvolver restrições e requisitos adicionais para reduzir o risco.

Às vezes, o mesmo perigo será gerado a partir das respostas a diferentes perguntas na lista criada. Ao final do processo, uma lista de cartões com diferentes perigos será organizada com os mesmos requisitos para que a equipe de desenvolvimento possa mitigar esses requisitos com o mínimo de esforço.

4 Comparação dos métodos de análise da segurança de sistemas

4.1 Análise para sistemas que não aplicam inteligência artificial

A análise da segurança e confiabilidade de sistemas críticos envolve extensa modelagem destes sistemas e/ou inúmeras informações estatísticas sobre o funcionamento do sistema, quando esse tipo de informação está disponível.

Cinco métodos de análise de segurança foram apresentados no capítulo 2, que podem ser definidos por algumas características. A maior parte dessa análise é realizada sob as observações de(ROUVROYE; V.D.BLIEK, 2002) com a adição da análise de STPA sobre outros métodos para sistemas não-IA. O método HAZOP é considerado tendo as mesmas características que a análise de especialistas.

Todos os métodos de análise de segurança utilizam as informações disponíveis sobre as falhas e probabilidades mais comuns do sistema em estudo para gerar uma análise. No entanto, o HAZOP exige a informação de um objetivo de segurança, ou seja, quão seguro o sistema deve ser para poder analisá-lo, e o FTA exige a definição da falha do sistema que será analisada em um determinado momento.

As etapas de execução dos métodos também diferem, o FMEA requer uma descrição completa do sistema, enquanto outros métodos são mais orientados para a descrição da interação entre as partes.

FMEA, HAZOP e STPA são métodos particularmente úteis para identificar situações potencialmente perigosas e fornecer à equipe de engenharia informações valiosas sobre o sistema a ser melhorado em seguida. Já as análises FTA e Markov apresentam abordagens mais quantitativas por meio da modelagem das falhas dos sistemas estudados pelas probabilidades de falhas nos subsistemas (FTA) ou pela probabilidade de transições de estado na análise de Markov.

As vantagens dos métodos mais quantitativos apresentados são a clareza dos resultados obtidos e a capacidade de comparação entre os diferentes modelos, que, em contrapartida, exigem maior complexidade e conhecimento prévio do comportamento dos subsistemas. Utilizando FMEA, HAZOP e STPA, mesmo sem o conhecimento prévio dos parâmetros de confiabilidade específicos de cada subsistema, é possível determinar cenários prováveis e úteis para aumentar a segurança. O STPA, em particular, demonstra uma

grande capacidade de identificar falhas resultantes de interações entre sistemas.

(RAJABLI et al., 2020), (ROUVROYE; V.D.BLIEK, 2002) e (WANG; CHUNG, 2022) não chegaram a uma conclusão sobre quais métodos são mais adequados para cada situação de engenharia, mas descrevem que, em termos de recursos, a análise de Markov aprimorada poderia fornecer mais informações sobre aspectos de segurança do sistema.

4.2 Análise relativa a sistemas que aplicam a IA

O objetivo desta parte é dar uma visão geral dos desafios colocados pela integração da inteligência artificial em sistemas de condução autônoma e quais os métodos a aplicar para garantir que o funcionamento destas funcionalidades seja seguro.

4.2.1 Algoritmos de aprendizado em veículos autônomos

Como a maioria das funções que integram a inteligência artificial utilizam Algoritmos de aprendizado de máquina*(Machine Learning)*, vamos nos concentrar apenas nesse aspecto da IA.

Os métodos de confiabilidade apresentados na seção 2, já estabelecidos na indústria, foram aplicados com sucesso em sistemas tradicionais, conforme descrito pela ISO 26262 e SOTIF. No entanto, esses padrões existentes não consideram as especificidades dos algoritmos de aprendizado, em particular as especificidades dos dados fornecidos para o aprendizado, da avaliação dos desempenhos, das incertezas, etc. Esta falta de padronização das análises necessárias representa um desafio para os fabricantes de veículos autônomos em termos de implementação de *Machine Learning* em seus sistemas.

O aprendizado de máquina é geralmente dividido em duas fases. Uma primeira fase de aprendizado e uma segunda de testes. A primeira fase consiste na estimação de um modelo a partir de dados escolhidos pelo desenvolvedor. Esta fase de aprendizagem consiste em resolver uma tarefa, como reconhecer um pedestre em uma imagem. A segunda fase consiste em testar o modelo definido anteriormente em relação à resolução das tarefas para as quais foi projetado. É importante notar que, mesmo durante a segunda fase, alguns sistemas conseguem continuar aprendendo.

Para cumprir os requisitos de segurança específicos da indústria automobilística, cada uma das duas fases anteriores deve ser objeto de muito cuidado durante as fases de implementação e de verificação.

4.2.2 Escolha do método de aprendizagem

Existem três métodos principais de aprendizado de máquina, que se distinguem pelas propriedades dos dados de entrada. Eles são :

- Aprendizagem supervisionada: as observações de entrada são rotuladas, ou seja, para cada dado de entrada uma classe, atributos, são associados. A intervenção humana é necessária para esta etapa. O sistema aprende a classificar as observações seguindo duas etapas. Um primeiro*(etapa de aprendizagem)*, durante o qual um modelo é determinado a partir das observações rotuladas. Então um segundo *(Etapa de teste)*, durante o qual as classes de outras observações são previstas;

- Aprendizado não supervisionado (ou clustering): os dados são fornecidos brutos (não rotulados), cabendo ao algoritmo classificar os dados de acordo com uma lógica que ele determinará;

- Aprendizado por reforço: o algoritmo, com base nas suas decisões prévias, otimiza suas decisões futuras.

4.2.3 Definição de critérios de aceitabilidade para funções usando Algoritmos de aprendizado

Existem inúmeras dúvidas sobre a segurança das funções de aprendizado de máquina. Para que um algoritmo de *Machine Learning* funcione corretamente, faz-se necessário que este seja concebido de forma apropriada a sua utilização, e que os bancos de dados que o formam também apresentem critérios adequados.

Por exemplo, para a construção de algoritmos de aprendizado de máquinas aplicados a sistemas de detecção de veículos autônomos, algumas questões de interesse levantadas em (WANG; CHUNG, 2022) e (BRABAND et al., 2020) devem ter suas respostas explicitadas:

- Quantos obstáculos devem ser detectados simultaneamente?

- Todos os objetos são do mesmo tipo/classe?

- Como verificar se os bancos de imagens de treinamento *(datasets)* são suficientemente diversificados e permitem que o algoritmo atinja um determinado nível de desempenho? (Tamanho do conjunto de dados? O que incluir no conjunto de dados?)

- Como definir o desempenho anterior a ser alcançado?

A avaliação de desempenho do algoritmo projetado para reconhecimento de imagem pode ser feita de acordo com diversas grandezas quantificáveis. Assim, os critérios de desempenho podem ser estabelecidos em relação a:

- Avaliação do número de falso-positivos e falso-negativos em um conjunto de dados de teste;

- Avaliação da taxa de observações corretamente classificadas;
- Avaliação da velocidade de execução da classificação;

É importante reduzir ao máximo o número de falsos negativos e falsos positivos detectados pelo algoritmo, pois um elemento sendo detectado como um falso positivo ou falso negativo representa um grande risco.

4.2.4 Escolha dos dados

As fraquezas que podem ser encontradas pelos algoritmos de *Machine Learning*, em termos de desempenho de detecção, são provenientes em primeiro lugar da qualidade dos conjuntos de dados usados.

Três conjuntos de dados distintos são usados para treinar um algoritmo de reconhecimento de padrões. O algoritmo, durante a fase de aprendizado, é treinado com um conjunto de dados de treinamento e um conjunto de dados de verificação. O objetivo do conjunto de dados de treinamento é criar um modelo de classificação. O conjunto de dados de validação, também utilizado durante a fase de aprendizado, visa verificar se o modelo é adequado, verificando a correta classificação das observações. O desempenho do algoritmo é então testado no final do aprendizado com um conjunto de dados de teste.

Os três conjuntos de dados devem ser escolhidos para cobrir uma grande parte dos eventos que serão provavelmente encontrados. Por exemplo, com imagens que representam pessoas em várias posições, em vários lugares da imagem, para diferentes brilhos, etc. As imagens positivas e negativas devem estar presentes em proporções semelhantes (com ou sem o alvo a ser detectado).

Outro desafio relacionado a dados é aprender a classificar observações respeitando rótulos predefinidos. De fato, uma observação pode ser caracterizada por vários atributos, tais como:

- Seu tipo: pedestre, veículo, bicicleta;
  - tipo de carro; presença de alguém de bicicleta;
- Seu tamanho, posição e distância;

Esses elementos também se aplicam ao reconhecimento do meio ambiente.

Um controle da rotulagem do conjunto de dados deve ser feito manualmente por um especialista designado. Erros de classificação comuns na detecção de alvos 3D são, por exemplo:

- Classificação incorreta;
- Limites de detecção mal posicionados (devido à oclusão parcial do alvo, brilho irregular, etc.);
- Fronteiras mal-dimensionadas

4.2.5 Arquitetura do programa

O *software* baseado em *Machine Learning* deve incorporar sistemas de proteção para garantir a operação segura do sistema. O primeiro deve conseguir lidar com comportamentos inesperados do algoritmo. Vários métodos são usados para garantir isso.

A supervisão pode ser adicionada, é caracterizada pela adição de um observador, um *software* adicional operando em paralelo com o algoritmo de *Machine Learning*, responsável por supervisionar seu comportamento com referências de segurança. Outras arquiteturas de *software* possibilitam aumentar a segurança, como redundância. Os resultados de vários algoritmos ML paralelos são usados para tirar uma conclusão. Quanto maior o número de MLs que dão o mesmo resultado, maior a probabilidade de que esses dados sejam os corretos.

As etapas de pré-processamento de entrada também são possíveis, como dimensionamento ou redimensionamento, amostragem, etc. a fim de aproveitar ao máximo as informações que eles podem fornecer.

4.2.6 Arquitetura do *Machine Learning*

Uma escolha criteriosa da arquitetura permite otimizar seu desempenho e velocidade de classificação, em particular devem ser escolhidas corretamente: as camadas, as funções de ativação, o tipo de camadas, etc.

Os algoritmos de Machine Learning são usualmente implementados na forma de redes neurais profundas (DNN).

5 Normas de segurança para veículos autônomos

Neste capítulo, as duas normas principais que regem a segurança de veículos são apresentadas: A ISO 26262, e a sua equivalente atualizada ISO 21448(ISO, 2011), (ISO, 2019).

As descrições à seguir são baseadas nos trabalhos desenvolvidos em (NASRI, 2018), (ISO, 2011), (ABDULKHALEQ et al., 2017) :

5.1 ISO 26262 - Segurança funcional

A Norma 26262, da Organização Internacional para Padronização, fornece uma abordagem baseada em risco específico automotivo para determinar o nível de integridade do sistema automotivo (ASIL).

5.1.1 Níveis ASIL e Hazard Analysis and Risk Assessment - HARA

A escala ASIL determina o nível de integridade ou risco de um determinado produto de acordo com as especificações da norma ISO 26262. Ela classifica esses níveis de integridade em quatro categorias: A, B, C e D. O nível ASIL de cada componente do veículo é determinado por uma análise HARA e cada nível corresponde respectivamente a uma colocação em uma escala crescente de periculosidade do sistema:

- ASIL A - O mais baixo
- ASIL B - Moderadamente baixo
- ASIL C - Moderadamente alto
- ASIL D - O mais alto

Para a realização da análise HARA, na ISO 26262, os seguintes termos são definidos:

- O Perigo é uma fonte potencial de danos causados por uma falha do sistema;
- Risco é a combinação da probabilidade de ocorrência do Dano e a gravidade deste;
- O elemento é um sistema ou uma função à qual a norma ISO 26262 é aplicada;

- A Análise de Perigos e Avaliação de Riscos (HARA) é um ”método de identificação e classificação de eventos Perigosos relacionados a um item conforme os níveis ASIL para evitar riscos excessivos”.
- O objetivo de segurança é o requisito de segurança de alto nível resultante do HARA.

HARA é um método específico da norma ISO 26262. Utilizado como meio de padronizar a análise das diversas fontes de Perigo e os níveis de risco, o HARA utiliza os métodos de análise de falhas citados anteriormente (como FTA, FMEA e HAZOP) para analisar a segurança do veículo.

O método HARA compreende duas análises distintas, a dos Perigos e a dos Riscos. A análise dos perigos identifica possíveis eventos não intencionais no caso de uma falha, enquanto a análise de risco se concentra na probabilidade, gravidade e controlabilidade das falhas do sistema.

A análise de risco, ou avaliação de risco, é descrita por meio destes três fatores:

- A exposição, ou a probabilidade de ocorrência de uma falha no componente analisado;
- A gravidade, ou a intensidade do dano que a falha no componente pode causar nas pessoas;
- Controlabilidade, ou a facilidade com que o motorista pode controlar o veículo em caso de falha.

Cada um dos três fatores é então colocado em uma escala de 1 a 3, sendo 1 o cenário mais favorável, E1 uma probabilidade muito baixa, S1 lesões leves e C1 uma situação facilmente controlada. E3 é uma possibilidade média, S3 incidentes com risco de vida e C3 uma situação de difícil controle do veículo.

Uma vez obtidos os parâmetros de exposição, gravidade e controlabilidade, sua combinação, incluindo o nível de Gestão de Qualidade,QM (Quality Management), dará o respectivo nível de integridade do sistema. Mostrado na tabela na Figura 2. Nessa figura, os três níveis de gravidade da falha são apresentados à esquerda(S1, S2 e S3), associados com a exposição da falha(E1 a E4) e a Controlabilidade do evento(C1, C2 e C3). O nível ASIL relativo a tal falha é apresentado na interseção entre as linhas e colunas.

Curiosamente, as normas ISO são dinâmicas e mudam constantemente, adaptando-se a novas possibilidades e requisitos de engenharia, como mostrou a inclusão de referências à cibersegurança na norma de 2018.

Atualmente, o principal ponto de discussão relacionado à norma ISO 26262 identificada por autores que estudam a segurança de veículos autônomos é que ela não

		Probability class	Controllability class		
			C1	C2	C3
Severity class	S1	E1	QM	QM	QM
		E2	QM	QM	QM
		E3	QM	QM	A
		E4	QM	A	B
	S2	E1	QM	QM	QM
		E2	QM	QM	A
		E3	QM	A	B
		E4	A	B	C
	S3	E1	QM	QM	A
		E2	QM	A	B
		E3	A	B	C
		E4	B	C	D

Figura 2 – Níveis ASIL da análise HARA(ISO, 2011)

contém uma análise de segurança para as interações entre sistemas, e que o fator humano não é considerado, quando na realidade, com carros autônomos, é muito provável que motoristas humanos sejam a causa de muitos dos acidentes enfrentados pelos sistemas desenvolvidos. (RAJABLI et al., 2020) afirma que “ainda não está claro se esse padrão pode ser usado efetivamente como em sistemas baseados em IA”.

Esta situação é uma das motivações para a criação da norma ISO 21448 e a tendência para a qual os veículos autônomos direcionam a indústria.

5.2 ISO 21448 - Segurança da Funcionalidade Pretendida - SOTIF

O conceito de ”funcionalidade pretendida”, *Safety of The Intended Functionality - SOTIF*, foi criado com base na inadequação da norma ISO 26262, e do conceito de ”Segurança Funcional”no projeto de sistemas elétricos baseados em grande volume de interações entre subsistemas e um volume muito grande de dados, como os sistemas de IA.

É por isso que o padrão ISO 21448 foi publicado em 2019, com o objetivo de ”Ser aplicado à funcionalidade pretendida onde a consciência situacional adequada é crítica para a segurança e onde essa consciência situacional é derivada de sensores complexos e algoritmos de processamento; especialmente sistemas de intervenção de emergência e Sistemas avançados de assistência à condução(ADAS) com níveis 1 e 2 nas escalas de automação J3016 da norma OICA/SAE.”(ISO, 2019).

5.3 Origem dos perigos na SOTIF

Existem quatro fontes principais de perigos analisadas no método HARA gerado com SOTIF. Seu relacionamento pode ser descrito através do diagrama da Figura 3, extraída da norma ISO21448.

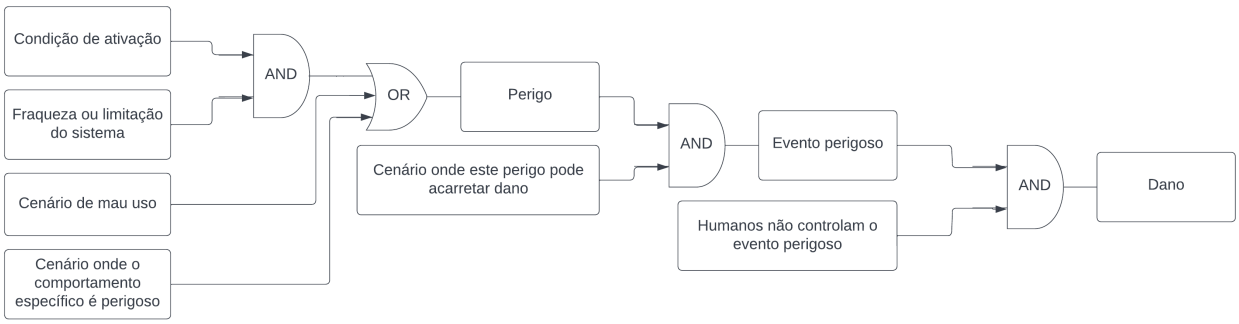

Figura 3 – Perigos e Riscos da norma ISO21448

Na figura 3, ficam definidos o perigo, o evento perigoso e a relação entre essas expressões e o dano, que é a ocorrência a evitar a todo custo. Na maioria das análises com veículos autônomos, o dano analisado é a ocorrência de uma colisão.

Se ocorrer uma condição de gatilho, levando diretamente a uma limitação do sistema, o veículo se tornará um perigo, ou seja, provavelmente criará uma situação suscetível de causar danos. O uso inadequado do veículo em cenários não previstos pela engenharia de segurança também pode tornar o veículo um perigo.

Um perigo, sob certas condições, torna-se um evento perigoso, ou seja, uma situação em que ocorrerão danos se nenhuma ação for tomada pelo operador humano.

5.3.1 Condições de Ativação

As condições de acionamento são condições ambientais que acionam alguma fraqueza ou limitação do sistema. Um exemplo de condição de ativação para um sistema de veículo autônomo pode ser, por exemplo, quando a câmera está obstruída ou uma fonte de luz forte próxima, como um raio ou feixe de luz, ilumina diretamente a câmera. É importante, em estudos preliminares, que uma equipe de especialistas liste todas as condições que podem levar a uma operação limitada do equipamento do veículo.

5.3.2 Fraquezas e limitações do sistema

Uma fraqueza ou limitação é um aspecto do sistema que é incapaz de operar com segurança na ocorrência de uma das condições de disparo mencionadas anteriormente.

Um desses pontos fracos pode ser o tempo que leva para o sistema recuperar a operação normal de uma câmera de detecção após um evento como brilho ou o tempo que leva para perceber que a imagem da câmera está obstruída.

5.3.3 Cenários de mau uso

Um cenário de mau uso, dentro do escopo da ISO 21448, é um cenário no qual o agente humano faz mau uso do sistema, resultando em comportamento inseguro do sistema. Tal situação pode ser simplesmente que o motorista perdeu a atenção e, não realizou as manobras esperadas, ou até mesmo forneceu manobras incorretas para a situação.

Uma pessoa envolvida em atividades ilegais, como dirigir sob o efeito de substâncias psicoativas ou mesmo ser apenas um motorista não qualificado, pode se tornar um cenário de mau uso.

5.3.4 Cenário onde o comportamento específico é perigoso

Este é um dos problemas mais difíceis na engenharia de segurança de veículos autônomos. Nesse tipo de cenário, o algoritmo do veículo autônomo realiza uma ação perigosa simplesmente porque o sistema não estava suficientemente preparado para essa situação, que geralmente se enquadra na categoria de situações perigosas e desconhecidas.

5.3.5 A relação entre ISO 26262 e ISO 21448

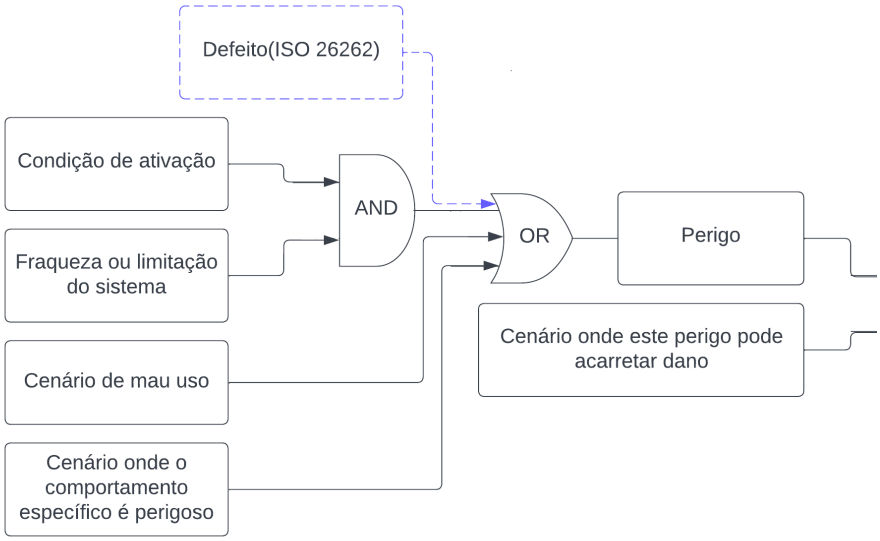

Figura 4 – Adição da definição de defeito da norma 26262 à SOTIF

A ISO 26262, conforme apresentado anteriormente, é uma certificação capaz de garantir, por meio estatístico, a segurança de sistemas elétricos e eletrônicos em caso de

mau funcionamento no âmbito da funcionalidade pretendida do sistema. A IA, no entanto, apresenta um novo conjunto de problemas devido à sua natureza de ”caixa preta”, ou seja, o próprio comportamento dos algoritmos de IA não é completamente compreendido pelas equipes de TI.

No entanto, a IA não substitui a importância da ISO 26262 e, de fato, como parte do processo de engenharia de segurança, agora é necessário separar as questões que precisam ser analisadas.

Uma maneira de visualizar facilmente esse novo relacionamento é adicionar um quarto elemento à instrução *OR* no diagrama mostrado na Figura 4.Este novo elemento representa o design da ISO 26262 sobre comportamento disfuncional, descrito pela análise HARA desta.

6 Estudo de caso: O Projeto PEGASUS

PEGASUS (Projeto para o estabelecimento de critérios, ferramentas e métodos geralmente aceitos, em conjunto com cenários e situações para a disponibilização de funções de direção altamente automatizadas) é um projeto financiado pelo Ministério Federal Alemão para Assuntos Econômicos e Clima que visa preencher as principais lacunas no teste e comissionamento de funções de direção altamente automatizadas(PEGASUS, b)(PEGASUS, a), A fim de garantir a introdução no mercado e a aceitação geral de veículos altamente automatizados (SAE nível 3+), o PEGASUS visa fornecer:

- Critérios de verificação de segurança da qualidade;
- Ferramentas e métodos geralmente aceitos por todos os Fabricantes Automotivos;
- Cenários e situações para comissionamento de funções de direção altamente automatizadas.

O objetivo é desenvolver um procedimento para testar as funções de condução automatizada, a fim de facilitar a rápida implementação da condução automatizada na prática.

6.1 O método PEGASUS

Os pontos a seguir visam descrever a arquitetura do método e como a informação evolui na cadeia de informações. Como demonstrado na Figura 5, o método articula-se em torno de 19 etapas destinadas a criar uma prova de verificação de segurança e uma última etapa em que é feita uma argumentação sobre a segurança do sistema de condução autônoma(PEGASUS, a).

O fluxo de informações passa por cinco blocos distintos de verificação e validação das funções do piloto automático. Esses blocos são interligados de tal forma que a informação que sai de um bloco é usada pelo próximo bloco. As principais etapas do método são:

- Definição de Objetivos;
- Processamento dos Dados;
- Estocagem e Tratamento dos dados em uma base de dados estruturada;
- Avaliação das funçõe de condução autônoma;

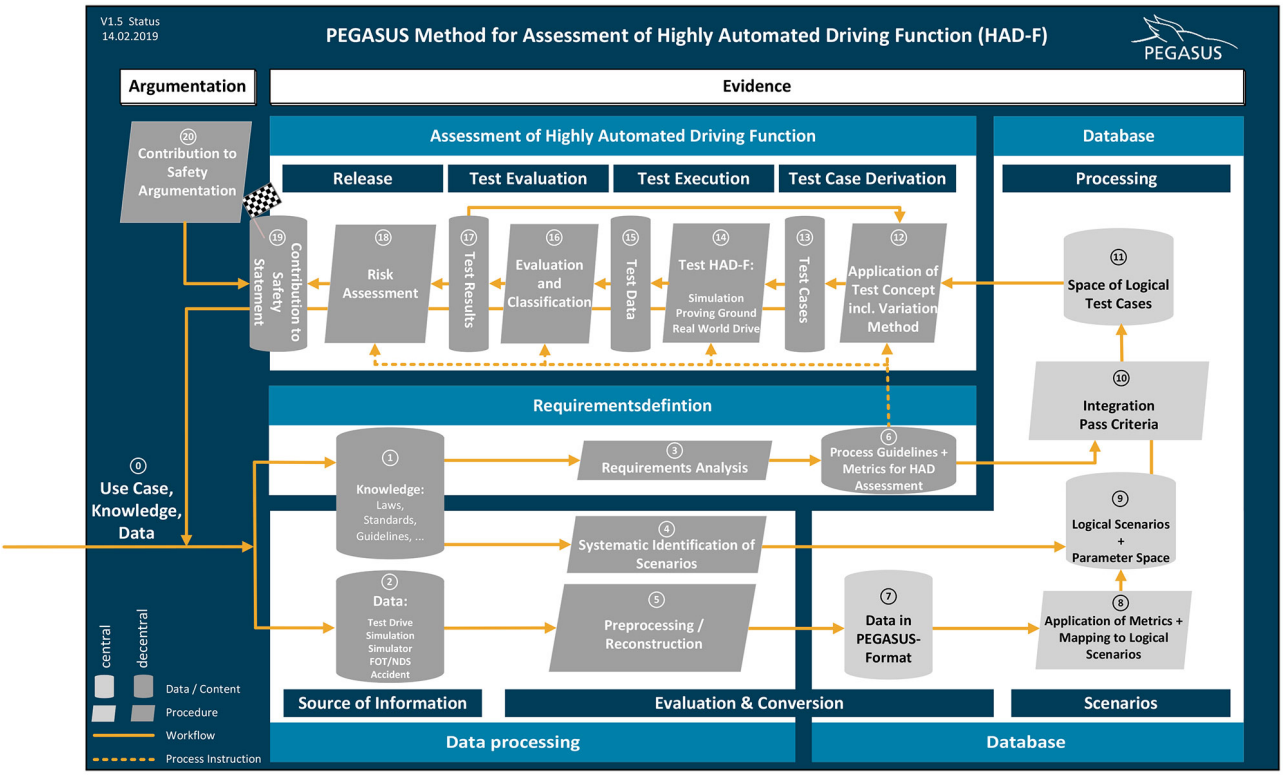

Figura 5 – Vista do conjunto do método PEGASUS(PEGASUS, a)

- Criação de um argumento sobre a segurança do sistema estudado;

6.2 Descrição detalhada do método PEGASUS

A partir da análise da Figura 5, o processo geral consiste no agrupamento de vários fluxos de dados. Como o método é aplicado linearmente, os dados que entram em um bloco vêm de um bloco anterior.

Antes de iniciar o processamento de qualquer informação, é importante no âmbito do método PEGASUS definir o assunto de sua análise e o cenário de sua utilização, aqui sendo respectivamente: as funções de condução altamente Autônoma de um veículo na rodovia, em que a limitação de dirigir na rodovia reduz o número de cenários que podem ser criados em comparação a uma direção em qualquer estrada.

Além disso, conforme observado no site do projeto, nem todos os cenários de condução em rodovias estão incluídos. Estão excluídos, por exemplo:

- Faixas de inserção e saída de rodovias;
- Situações de condução com condições meteorológicas extremas (nevoeiro espesso, estrada muito escorregadia);

- Modificações temporárias de estradas (locais de trabalho).

Esses cenários são excluídos em favor de cenários mais frequentes:

- Trânsito intenso e lentidão;
- Mudança de faixa por outros veículos;
- Velocidade do carro autônomo entre 0-130 km/h:
- Observância da sinalização.

Para entender o funcionamento e os pontos fortes do método, será feita uma descrição dos principais blocos, explicando os conceitos e tipos de dados envolvidos.

6.2.1 Processamento de dados (blocos 1/2/4/5)

O método PEGASUS inicia-se com uma etapa de coleta de informações referentes ao caso do motorista de rodovia *(highway chauffeur)*, como os dados de "conhecimento"sobre normas e leis de condução em autoestradas (ISO 26262 e SOTIF para segurança, leis de infraestrutura rodoviária, etc.), mas também dados de condução de várias fontes (condução real, em simulador, etc.).

As duas primeiras etapas de processamento de dados são feitas de forma independente, por um lado os dados de condução são usados para serem colocados em um formato que permitirá que sejam usados em cada etapa do método, por outro lado, os dados de*knowledge* serão usados para configurar um método para identificar cenários de direção. Importante observar que a introdução de um novo elemento no bloco *knowledge*, como uma nova função de automação, diversifica os cenários que podem ser encontrados, à medida que novos recursos são introduzidos.

A identificação sistemática de cenários perigosos possibilita a criação de um banco de dados de cenários a partir do qual serão definidas e avaliadas as funções do piloto automático. O objetivo é que as funções de direção autônoma sejam seguras em situações propensas a colisões. Esses cenários são armazenados em um banco de dados e podem ser representados na forma de uma árvore, onde cada vértice corresponde a uma situação de perigo, e a passagem de um vértice para outro é feita adicionando um novo elemento ao cenário inicial. Desta forma é possível criar uma base razoavelmente representativa de cenários a serem testados.

6.2.2 Definição das exigências (blocos 1/3/6)

Quanto à parte de processamento de dados, as diversas informações do bloco *knowledge* são usados, mas para um propósito diferente. Nesta parte são definidos os

requisitos em torno do desempenho do sistema de condução altamente automatizado. Baseiam-se em critérios como aceitação pelo público em geral, quantificação do risco aceitável associado ao piloto automático, etc. Esses critérios serão então usados como uma diretriz para a avaliação de segurança.

6.2.3 Base de dados (blocos 7 a 11)

Os dados dos dois processos anteriores são então armazenados em bancos de dados. Esses bancos de dados preenchem a lacuna entre dados brutos e requisitos e testes. Os dados no formato PEGASUS são usados para definir cenários lógicos: cenários lógicos são representações do ambiente em torno do objeto de observação associado a grandezas físicas. Esses cenários lógicos são formados a partir de 6 camadas principais:

- Camada 1: descrevendo a estrada: por sua geometria, seus limites;
- Camada 2: descrição da sinalização: barreiras levantadas, sinalização;
- Camada 3: descrevendo modificações temporárias de trilha;
- Camada 4: descrevendo os objetos dinâmicos ao redor do veículo observado;
- Camada 5: descrição do ambiente (brilho, tempo);
- Camada 6: descreve a troca de informações com o veículo.

Em segundo lugar, as condições de sucesso ou falha estão associadas aos cenários lógicos. Esses resultados são definidos conforme as métricas contidas no bloco 6*Process Guidelines and metrics for HAD assessment*, e nas informações já geradas no bloco 9. A integração dos critérios de sucesso de um cenário lógico permite definir um caso de teste. A saída de informação do bloco 10 representa a verificação da relevância das configurações nas quais as funções de condução automatizada serão testadas, que é integrada ao espaço de testes no bloco 11.

A partir desta fase é possível realizar uma verificação da segurança do sistema de condução altamente autônomo.

6.2.4 Avaliação da função de condução altamente autônoma (blocos-12 a 19)

Os dados de entrada para esta seção são os requisitos de segurança definidos no bloco 6, bem como os testes lógicos previamente definidos no bloco 11. Os principais objetivos desta seção são a execução e avaliação de testes. A avaliação é feita por três métodos: em simulador, em local de teste ou em condução real. A complexidade e baixa

frequência de ocorrência de determinados cenários exige, por exemplo, verificação em estrada, em condições reais, enquanto cenários pouco relevantes e frequentes requerem apenas testes em campo de provas. A reprovação em um teste requer modificações no veículo para que o teste seja aprovado na próxima tentativa. A execução dos testes permite, em função dos requisitos, quantificar a segurança das funções de condução Autônoma.

6.2.5 Argumento sobre segurança antes da implantação em larga escala (bloco-20)

A etapa final do método PEGASUS consiste em um argumento da segurança do sistema em relação aos resultados das partes anteriores. São considerados os resultados anteriores, verificando-se a sua pertinência, consistência, formalização, etc. a fim de verificar o cumprimento do nível 3 SAE definido. A utilização de 5 etapas distintas permite que o processo que garanta uma argumentação geral coerente e ajustável.

6.3 Percepções sobre o método

6.3.1 Forças e fraquezas

Em primeiro lugar, a própria essência da verificação das funções de condução por uma abordagem de cenário permite uma grande flexibilidade. Adaptando os axiomas iniciais, será possível eventualmente passar da condução em autoestrada para a condução em ambiente urbano. Neste caso, o caso de uso mudaria e, consequentemente, os dados de condução de entrada, bem como os padrões a serem considerados.

A utilização de bases de dados é propícia a uma utilização comum do método pelos vários intervenientes na indústria automotiva. Armazenar e compartilhar dados de condução, cenários e resultados permite uma progressão constante das técnicas de verificação. Por isso, em sua iniciativa, as equipes do projeto PEGASUS recomendam acordos de cooperação entre fabricantes.

Por outro lado, este método deve ser acoplado a métodos complementares, deficiências relacionadas à IA, processamento de informações recebidas pelo veículo, deficiências relacionadas a sistemas de visão, etc. não são considerados. Certas condições de operação do veículo, como condução em condições climáticas extremas e em áreas em construção são excluídas dos cenários, consequentemente a operação do veículo não é verificada lá.

7 Adição às estratégias existentes

É definido em (DEEL Certification Workgroup, 2021) que os problemas de segurança de sistemas de segurança crítica que utilizam machine learning, tais como veículos autônomos, emergem de três incertezas:

- A incerteza dos sensores e o algoritmo de percepção
- A incerteza do algoritmo de planejamento
- A incerteza do algoritmo de controle

7.1 A incerteza dos sensores e o algoritmo de percepção

Muitos estudos foram realizados sobre a segurança e garantia da operação de sistemas de IA, como mostra o exemplo a seguir (RAJABLI et al., 2020), essas validações fornecem estatísticas úteis sobre o desempenho desses sistemas, em particular para o reconhecimento de padrões que é usado principalmente para a percepção do mundo no veículo autônomo graças às diferentes tecnologias utilizadas (LIDAR, câmeras de vídeo, Sensores infravermelhos). No entanto, um problema de segurança persistente com esses algoritmos é que eles são difíceis de testar para os chamados ambientes de ”mundo aberto”, e suas principais aplicações para sistemas de missão crítica estão sob variabilidade de entrada muito menor do que o que é visto em Veículos Autônomos médios.

Esses algoritmos utilizados para a percepção do ambiente sofrem de um problema de viés, trazido por seus bancos de dados. Esse problema de viés pode ser mitigado por uma boa seleção de dados para treinamento desses sistemas, mas nunca pode ser completamente eliminado, sendo especialmente difícil para as agências de segurança de todo o mundo saber como testar adequadamente esses novos sistemas.

Além disso, graças aos algoritmos de verificação de IA, é possível alcançar o nível de segurança dessa IA sem necessariamente considerar os problemas físicos do sistema de aquisição. Existem muitos problemas que podem ocorrer com o sensor que podem interferir no algoritmo inesperadamente, como muita ou pouca luz no ou nos sensores utilizados, fazendo com que ele envie informações incorretas pré-processadas para o algoritmo de reconhecimento. o algoritmo de tomada de decisão, etc. STPA talvez possa ser usado para minimizar tais problemas, aplicando uma verificação de segurança ao subsistema que consiste na ”câmera”e no algoritmo de reconhecimento.

7.2 A incerteza do algoritmo de planejamento

Este tópico é sobre Ações de Controle Inseguras e como evitá-las e gerenciá-las. STPA é uma abordagem que parece particularmente adequada para gerenciar interações entre o veículo autônomo e outros veículos na estrada, bem como para gerenciar problemas decorrentes de falhas não identificadas do sistema de aquisição e para criar restrições suficientemente seguras sobre os atuadores.

7.3 A incerteza do algoritmo de controle

O algoritmo de controle é responsável por transformar as ações de alto nível pretendidas, como aceleração, ângulo de giro, em ações de baixo nível para os atuadores, na forma de entradas elétricas.

Quando se trata da segurança dessas incertezas, os algoritmos podem ser limitados com relativa facilidade pela utilização de restrições e requisitos aplicados aos algoritmos de aprendizado de máquina, e podem ser aplicadas tanto por meio do processo de reforço e aprendizagem das mesmas quanto à posteriori, pela utilização de sistemas específicos.

7.4 Sugestão para a melhoria da validação e verificação utilizando dados reais

No método Pegasus (PEGASUS, b), na Seção 12, ”Aplicando o Conceito de Teste”, afirma-se que para avaliar a criticidade de diferentes cenários e descobrir os mais relevantes, alguma medida de severidade deve ser obtida por variação de parâmetros.

Afirmou-se que esses parâmetros deveriam ser relativamente simples para manter os custos computacionais baixos e permitir uma maior variação nos parâmetros do programa. Este problema é uma realidade para todas as técnicas de Verificação e Validação de segurança por simulação.

Dois dos parâmetros utilizados para análise de simulação e avaliação de gravidade pelo método PEGASUS foram o ”Tempo Para Colisão estimado”(TTC) e o parâmetro mais complexo ”Tempo Para Colisão Adicionado pela Velocidade da Colisão”(TTCVCOL). Como esses valores são usados para avaliar o risco de cada cenário, quanto mais negativos forem os valores, maior será o risco. O TTC não apresenta de imediato uma medida de gravidade, exigindo uma avaliação mais aprofundada do cenário que originou a situação crítica. Ambas as métricas são visíveis na figura 6.

Estes gráficos foram obtidos pelos arquivos descritivos do método PEGASUS (PEGASUS, a), e ajudam a descrever a utilização de simulações para buscar cenários

perigosos.

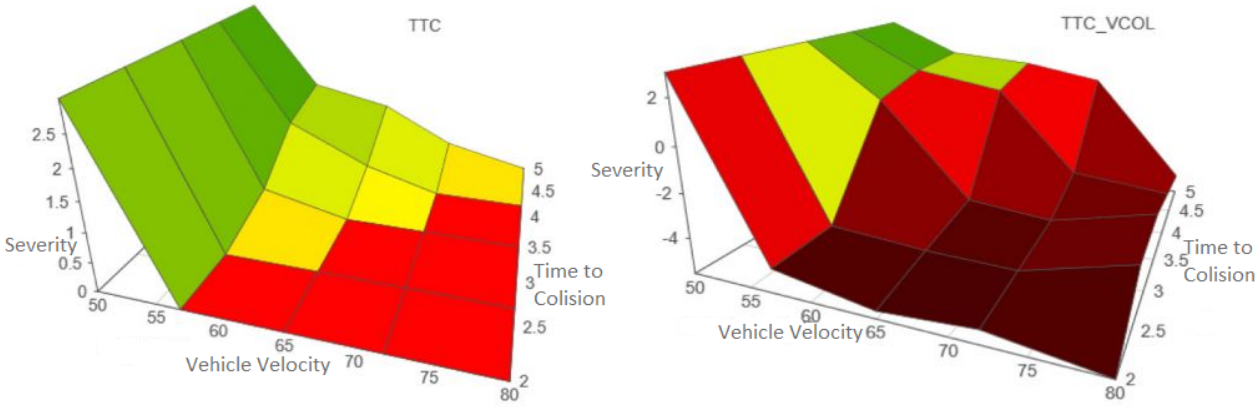

Figura 6 – Aplicação dos conceitos de teste de TTC e TTCVCOL (PEGASUS, b)

Em (PEGASUS, b), na seção 12, a discussão é seguida por 5 perguntas/problemas para descrever melhor o desafio da avaliação de segurança baseada em cenários:

- Como são definidas as métricas que fornecem uma medida realista das violações de segurança, incluindo a gravidade do incidente? Além disso, o cálculo dos modelos métricos não deve ser muito caro (por exemplo, uma simulação de colisão de elementos finitos provavelmente não seria aceitável devido ao grande número de simulações de cenários necessárias).
- Como definir métricas que permitam algoritmos de otimização mais eficientes?
- Como definir parâmetros que não expõem vários mínimos locais espúrios ou grandes ”áreas de platô”?
- Como dimensionar várias métricas para obter valores comparáveis, caso várias métricas devam ser usadas em um cenário lógico?
- Considerar que efeitos de amostragem e erros numéricos em simuladores podem dificultar a identificação de um cruzamento zero (por exemplo, para TTC).

A ideia a seguir é voltada para resolver os dois primeiros problemas com Validação e Verificação baseada em cenários:

A detecção em AVs estará sujeita a medidas de desempenho probabilísticas em suas diferentes condições de operação. Uma dessas medidas poderia ser a probabilidade de identificação de um objeto a partir de um certo número de pontos, dependendo das

Table 1 Effective range and uncertainty of Velodyne lidar under different weather conditions

Effective range (ER) in various conditions and thresholds (m)								
Point threshold	Clear Day	Uncertainty in measuring ER on clear days (%)	Rain	Uncertainty in measuring ER on rainy days (%)	Snow	Uncertainty in measuring ER on snow days (%)	Fog	Uncertainty in measuring ER on foggy days (%)
5	99.9	2.1	91.8	2.8	96.0	2.6	56.8	2.3
10	74.8	2.1	70.3	2.8	72.8	2.6	45.8	2.3
20	56.0	2.2	53.9	2.9	55.2	2.8	36.9	2.5
40	41.9	2.3	39.2	2.9	40.9	2.9	29.8	2.5
80	31.4	2.3	30.6	3.1	31.0	2.9	22.0	2.7
200	21.4	2.4	20.2	3.2	21.0	3.0	16.1	2.9

Tabela 1 – Verificação de alcance efetivo para LIDAR (ABDO et al., 2022)

capacidades do LIDAR/sensor, conforme estudado em (ABDO et al., 2022), cujos resultados podem ser vistos na Tabela 1.

Propõe-se que essas probabilidades possam então ser implementadas em Validação e Verificação baseada em cenários diretamente no modelo, por meio de um deslocamento ou multiplicação da função de risco. Usando o exemplo da Figura 1, se o número de pontos necessários para ter uma probabilidade suficiente de detectar um veículo que está chegando é conhecido, então as probabilidades para as diferentes condições climáticas podem ser usadas para aumentar a sensibilidade do realismo e as informações do modelo.

8 Conclusão

Os requisitos da indústria e de segurança têm evoluído constantemente no que diz respeito aos sistemas baseados em inteligência artificial, principalmente os sistemas sensoriais e de tomada de decisão, pois é esperado que esses sistemas às vezes encontrem situações inesperadas e que seja impossível testá-los em todas as situações. No entanto, isso não significa que esses sistemas sejam necessariamente perigosos, mas que os métodos e especificações de segurança devem ser atualizados para acompanhar essas mudanças tecnológicas.

Através deste trabalho, foi possível observar que os métodos clássicos de avaliação de segurança continuam relevantes e importantes para verificar a segurança dos veículos autônomos. Porém, ficou evidente que as validações de segurança destes sistemas necessitarão sucessivamente de mais dados e simulações para validar um conjunto de cenários estatisticamente relevantes, além de análises sistemáticas como as fornecidas pela metodologia STPA.

Como uma das metas deste trabalho, foi apresentada uma proposta de utilização de dados de sensores na modelização de cenários para a verificação e validação de veículos autônomos tomando por base os desenvolvimentos feitos pela metodologia PEGASUS.

Finalmente, os avanços em confiabilidade apresentados neste trabalho apontam para que a utilização dos veículos autônomos em um nível SAE 4 e superior seja regulamentada nos próximos anos, dado condições suficientes para comprovar sua segurança sejam desenvolvidas, notadamente graças ao progresso no uso de simulações para Validação e Verificação algorítmicas como foi feito no projeto PEGASUS. No estado atual da pesquisa de segurança, é provável que veículos autônomos possam operar com total autonomia em situações restritas primeiro, como rodovias ou cidades com níveis de infraestrutura necessários.

Referências

ABDO, J. et al. Effective Range Assessment of Lidar Imaging Systems for Autonomous Vehicles Under Adverse Weather Conditions With Stationary Vehicles. *Journal of Risk and Uncertainty in Engineering Systems, Part B*, set. 2022. Vol 8. Citado na página 39.

ABDULKHALEQ, A. et al. Using STPA in Compliance with ISO 26262 for Developing a Safe Architecture for Fully Automated Vehicles. *Automotive - Safety Security 2017*, p. 149–162, 2017. Citado 2 vezes nas páginas 18 e 25.

American National Standards Institute. *ANSI/UL 4600: Standard for Safety for the Evaluation of Autonomous Products*. [S.l.], 2020. Citado na página 12.

AULINAS, J.; SJAFRIE, H. *AI for Cars*. [S.l.]: Chapman and Hall/CRC, 2021. ISBN 9781003099512. Citado na página 13.

BRABAND, J. et al. On Safety Assessment of Artificial Intelligence. *Dependability Journal*, 2020. Vol. 20 No. 4. Citado 4 vezes nas páginas 10, 12, 13 e 22.

DEEL Certification Workgroup. Machine Learning in Certified Systems. mar. 2021. Citado na página 36.

International Organization for Standardization. *ISO 26262: Road Vehicles – Functional Safety*. 2011. Citado 5 vezes nas páginas 5, 12, 18, 25 e 27.

International Organization for Standardization. *ISO 21448: Road Vehicles – Safety of the Intended Functionality*. 2019. Citado 2 vezes nas páginas 25 e 27.

ISHIMATSU, T. et al. Modeling and Hazard Analysis using STPA. IAASS, 2010. Citado na página 18.

MAHAJAN, H. S. et al. Application of STPA to a lane-keeping Assist System. *Reliability Engineering and System Safety 167*, p. 177–183, 2017. Citado na página 18.

NASRI, S. A. Application of STPA methodology to an automotive system in compliance with ISO26262. 2018. Trabalho de Conclusão de curso, University of Stuttgart. Citado na página 25.

O'KANE, S. *Tesla can change so much with over-the-air updates that it's messing with some owners' heads*. 2018. The Verge. Disponível em: <https://www.theverge.com/2018/6/2/17413732/tesla-over-the-air-software-updates-brakes>. Citado na página 14.

PEGASUS. *Pegasus Method*. Main Website for project description and information. Disponível em: <https://www.pegasusprojekt.de/en/pegasus-method>. Citado 4 vezes nas páginas 5, 31, 32 e 37.

PEGASUS. *Pegasus Method: An Overview*. Disponível em: <https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf>. Citado 4 vezes nas páginas 5, 31, 37 e 38.

RAJABLI, N. et al. Software verification and validation of safe autonomous cars: a systematic literature review. *IEEE Access*, 2020. Citado 5 vezes nas páginas 9, 15, 21, 27 e 36.

ROUVROYE, J.; V.D.BLIEK, E. Comparing Safety Analysis Techniques. *Reliability Engineering and System Safety 75*, p. 289–294, 2002. Citado 2 vezes nas páginas 20 e 21.

Society of Automotive Engineers. *Levels of Driving Automation*. 2014. Disponível em: <https://www.sae.org/blog/sae-j3016-update>. Citado 2 vezes nas páginas 5 e 9.

WANG, Y.; CHUNG, S. H. Artificial Intelligence in safety-Critical Systems: A Systematic Review. *Industrial Management & Data Systems*, 2022. Vol. 122 No.2. Citado 3 vezes nas páginas 12, 21 e 22.

ZARKADAKIS, G. *5 Things AI can do better than humans*. 2019. Disponível em: <https://georgezarkadakis.medium.com/5-things-ai-can-do-better-than-humans-6dacad065848>. Citado na página 11.