Explain It for Safety: Explanations for Risk Mitigation

Maike Schwammberger[1][0000−0002−3344−6282], Astrid Rakow[2], Lina Putze[2][0000−0002−7443−1191], and Akhila Bairy[1][0000−0002−8796−1474]

[1] Karlsruhe Institute of Technology, Karlsruhe 76131, Germany {schwammberger, akhila.bairy}@kit.edu mase.kastel.kit.edu
[2] Inst. of Systems Engineering for Future Mobility, German Aerospace Center (DLR) e.V., Oldenburg 26121, Germany {astrid.rakow, lina.putze}@dlr.de

Abstract. Traffic conflicts are critical scenarios where autonomous vehicles (AVs) must navigate hazards while prioritising safety. These situations often involve significant risks for all traffic participants. We postulate a key advantage of incorporating explanations in such high-risk scenarios: when delivered to the right recipient at the right time, an explanation can reduce risks in conflict situations. This paper advocates for explanations that not only enhance understanding of conflicts but actively contribute to their resolution. We illustrate our case with examples of high-risk conflict scenarios and derive essential design processes and requirements for risk-mitigating explanations.

Keywords: Autonomous Traffic Agents · Conflicts · Explainability · Risk Mitigation · Safety · Game Theory

1 Introduction

When designing autonomous vehicles (AVs), dealing with *conflict situations* must be part of a structured AV system design process. Conflicts comprise traffic situations, in which traffic rules or central safety goals of an AV are violated [16]. Violated safety goals could include injuries or fatalities amongst vulnerable road users (VRUs).Consider the example sketched in Fig. 1. *The AV A is facing an obstacle on the road. Three options are available for A to handle this conflict situation: (a) to brake and collide with the obstacle, (b) to drive into the opposing traffic where another AV B is approaching or (c) whether to drive into the road border.* Clearly, all decision alternatives entail a high *risk* to violate important safety goals of the AV. These risks could include injuries or fatalities of vulnerable road users (VRUs) [20]. For instance, if the AV *B* does not perceive *A*, a collision might be unavoidable. On the other hand, driving into the obstacle would also lead to a highly risky, and less predictable, situation, while crashing into the dense reflector posts at the road side certainly means fatal damage.

This example illustrates the close connection of the concepts of *conflict* and *risk* that are at the centre of our contribution. Clearly, conflict situations that entail an unacceptable risk must be mitigated by suitable safety mechanisms to such an extent that the risk becomes tolerable. This paper is specifically concerned with hazards that arise because multiple agents are in conflict.

Our key contribution is to discuss the role of *explainability* as a mechanism to reduce risk in conflict situations. *In the sketched example, AVs A and B need to coordinate their behaviour to lower the risk that is associated with the situation.* We argue that incorporating explainability into such conflict situations is an essential safety mechanism for resolving conflicts and mitigating associated risks. This paper describes a methodical approach to derive *requirements* for explanations that are fit for risk mitigation. Such requirements would, for instance, contain information on when to explain what and to whom.

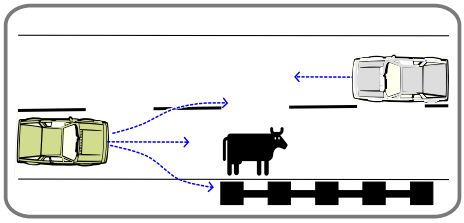


Fig. 1: Obstacle ahead example: The green AV A to the left has to decide between high-risk manoeuvres due to a sudden obstacle on the road. The *other* car B is approaching on the opposite lane.

For explanations in conflict situations, we take two types of explanation recipients into account: AVs and human agents (HA). It must be noted that explanations to other AVs are much more complex to design than simple V2V communication messages: The explanation content must be tailored towards the specific situation, and towards AVs with different capabilities. Potential human agents in need of explanations can be an AV driver that must take-over control, an engineer that intends to optimise the system, or a vulnerable road user (e.g. a pedestrian or bicyclist). For human agents, explainability design processes vary greatly [34]. In particular, *human models*, are difficult to define, but necessary in order to integrate them into safety mechanisms that are supported by tools and formal methods [6, 57].

Our contribution is structured as follows: we discuss terminology and related work for our three central notions of *conflict*, *risk* and *explanation* in Sect. 2. We motivate the need for explanations in risk mitigation in Sect. 3 and derive crucial explainability requirements for system design processes in Sect. 4. In Sect. 5, we integrate the derived explainability requirements into run-time explainability processes. We discuss the specific challenges of explaining to humans in Sect. 6 and conclude this paper in Sect. 7.

2 About Conflicts, Risks and Explanations

Our key contribution is to promote explanations to be used for minimising risk in conflict situations. To that end, we provide a brief overview over the notions of *conflict*, *risk*, and *explanation* in this section. We refer to respective definitions that we base our arguments on, and we briefly discuss the current research landscape around the respective terms.

2.1 Conflict

Conflicts of autonomous traffic agents (ATAs) have been extensively studied from various perspectives, including collision avoidance, resource competition, and ethical dilemmas [26, 35, 52]. Traditionally, conflicts have been considered primarily as collisions or accidents in the field of transportation [43, 10], focusing on safe navigation and path planning of individual agents [45, 21]. With the advancements in autonomous systems, the focus expanded to conflicts arising from agent interactions and the need for global coordination and negotiation strategies [48, 12]. With the advent of autonomous vehicles and intelligent transportation systems, conflicts also include *ethical conflicts* (e.g. [26, 35]), *legal conflicts* (e.g. [19, 31]) and conflicts due to human-agent interaction e.g. [27].

In this paper we focus on the conflict notion of Damm, Fränzle et al. [11]. In [11] the authors study conflict resolution mechanisms of ATAs using a general conflict model that frames conflict situations game-theoretically. This model is based on Galtung's conflict triangle [17]. Game theory is a research

field that has been developed by von Neumann and Morgenstern [42] and others in the mid-20th century. It studies interactive decision-making where the outcome for each player depends on the actions of the other players [40]. Game theory has been successfully applied to various domains such as economics, political science and computer science.

Let us first assume complete information. Following Galtung, a conflict can be captured between two agents in a game-theoretic setting as follows: Let two agents A and B have prioritised lists of goals G_A, resp. G_B. Agents A and B are in conflict, if

($C1$) *there is a winning strategy for A if A would control B, and*
($C2$) *There is a winning strategy for B if B would control A, and*
($C3$) *there is no strategy to achieve both the goals of A and B, $G_A \cup G_B$.*

$C1$-$C3$ describe that any strategy *(i.e. rational course of actions)* of A to achieve its goals is opposing the strategies of B to achieve its goals. In such a situation, A cannot decide on a strategy independently: it needs the help of B, but B is not inclined to help – and vice versa B cannot decide independently. *In the scenario of Fig. 1, A intends to drive onto the lane for oncoming traffic to avoid both the obstacle on the road, g_c, and the crash barrier at the road side, g_b. Suppose that B's lane has uneven and slippery surfaces near the road border. Since B wants to keep its safety goal, g_r, and keep its speed, g_s, it is not inclined to move aside onto the road border. It is hence not possible to realise all goals of A and B,* $\bigcup_{x\in\{c,b,r,s\}}\{g_x\}$.

The notion of conflict from [11] is an epistemic notion that studies conflicts from the viewpoint of an agent. Both, incomplete information *("What game is played?")*, as well as imperfect information *("What is the current state?", "What actions have been played?")*, are possible. An agent A perceives the real world only via its observations from which it constructs a belief about the situation. We assume A to be rational, meaning that it always tries to find a winning strategy to achieve its goals G_A. However, A can only evaluate whether a strategy is appropriate according to its beliefs. Wrong beliefs due to e.g. misperception may also cause A to wrongly believe that there is a conflict. *In our example, B may only think, that its lane border is slippery while in fact it is not. It hence wrongly believes to be in a conflict with A, when asked to move aside.*

We say that a conflict between A and B is *resolved* for A, if A knows what strategy to choose. That is from A's point of view it has a winning strategy. In [11], an approach to conflict resolution is presented that reveals increasingly more information – thus changing A's assessment of the situation – until finally a negotiation phase is started where A and B have to agree which of their goals can be sacrificed. The negotiation itself is out of the scope of [11].

2.2 Risk

In compliance with the EU implementing regulation for the admission of AVs, manufactures of AVs are obligated to provide a safety concept accompanied by a documented argumentation demonstrating that their system does not cause any unreasonable risk. This means that there arise no risks that are considered unacceptable based on valid societal moral principles [14]. In the context of road vehicles, *risk* is defined as "combination of the probability of occurrence of harm and the severity of harm" [25, 24], whereas *harm* refers to "physical injury or damage to the health of persons" [24].

In order to build a safety concept and a corresponding safety case that provides a comprehensive and reasonable safety argument supported by corresponding evidences, a systematic safety process has to be integrated in the system's design process [55]. Such a safety process has to be compliant with ISO 26262 and ISO 21448 covering both *functional safety* and *safety of the intended functionality (SOTIF)* [24, 23]. The former is concerned with risks resulting from failures or unintended behaviour of electronic or electrical systems [24]. The latter deals with the inherent safety of the specification

itself, i.e. risks that are caused by insufficiency of the specification or the performance, like limited technical capabilities or gaps [23].

An important part of a safety process is a systematic *hazard analysis and risk assessment (HARA)* that identifies and evaluates hazardous events in which a specific system behaviour of the system (hazardous behaviour), resulting from system failures or functional insufficiency, may provoke harm, as e.g. proposed by [33]. The traffic situation that is depicted in Fig. 1 illustrates such a hazardous event. In this contribution, we primarily encounter two types of risks: the risk of physical injuries to the passengers of the ego vehicle and the risk of physical injuries to the passengers of the oncoming vehicle. We perceive three behaviour options for the ego vehicle in this example: (a) an evasive manoeuvre onto the lane for oncoming traffic, (b) lane keeping with braking, and (c) an evasive manoeuvre towards the road border. Each of these options provoke one or both of the mentioned risks, thereby constituting hazardous behaviours.

As part of the HARA, safety goals and corresponding acceptance criteria have to be defined for each hazardous event. These safety goals serve as a foundation to derive safety requirements for the different functionalities and technical capabilities. Assuming that a detailed analysis of a hazardous event reveals that the associated risk exceeds an acceptable level, the implementation of safety measures becomes mandatory. The safety measures may address either the failures and functional insufficiency directly, thereby reducing the probability that a hazardous behaviour occurs, or they may aim at enhancing the controllability of the situation or a mitigation of the potential severity. During the verification and validation process it has to be ensured that the safety measures are efficient in the sense that they reduce the residual risk to an acceptable level.

2.3 Explanation

The Cambridge dictionary defines an explanation as "the details or reasons that someone gives to make something clear or easy to understand"[3]. Besides improved understandability, explanations are often linked to an increase in trust into automated systems [37, 15]. In the following, we introduce explainability terminology, mainly focusing on definitions that have been introduced in Köhl et al. [29] and Chazette et al. [9], adapting them towards the definitions that we have used before in [51]. The following elements are essential for engineering an explanation:

(i) ***explananda*** X , or "phenomena", of the system of interest (singular: explanandum),
(ii) ***context*** C for defining the environmental situation that influences the systems' behaviour, and the observed explananda,
(iii) ***stakeholder group*** G as explanation recipients,
(iv) ***goal*** Θ of an explanation for a stakeholder group G, and
(v) ***means*** M for producing the explanation.

Reconsider our example from Fig. 1; (i) an *explanandum* X could be the automated distance-keeping functionality of AV A, noticing that the distance to the obstacle is too small, (ii) a *context* C is the specific traffic situation that is depicted, (iii) a *stakeholder group* G could be passengers of AV A or another AV B, (iv) the explanation *goal* Θ could comprise that the group G can trust the system more after the explanation or that AV B is aware of A's conflict, and (v) the *means* M for producing an explanation can be a voice assistant in case of a human stakeholder group G or a vehicle-2-vehicle message in case of a non-human explanation recipient.

This terminology allows for a more formal definition of what an *explanation* is.

[3] `https://dictionary.cambridge.org/de/worterbuch/englisch/explanation#google_vignette` (accessed on 19.12.2024)

Definition 1 (Explanation [51]). *An explanation E for a given explanandum X and a target group G of stakeholders with an explainability goal Θ is a piece of information (or evidence) that makes the explanandum X understandable by G with respect to the goal Θ.*

An exemplary explanation E could comprise that the automated car shows a visual, explaining to the group G of passengers that driving onto the lane for oncoming traffic will be necessary to avoid a collision with the obstacle in front (the latter one describing the context C). In this case, the explanation goal Θ could be that the explanation recipients in G can trust the AVs decision better. Following [29], a system S can be named "explainable", iff in any context C, for an explanandum X and stakeholders G, an explanation E can be provided. It can further on be called *self-explainable*, if it is capable of deriving E without the aid of an entity outside of S.

The general need for integrating self-explainability capabilities into system engineering processes of AVs can be substantiated by the IEEE Standard for Transparency of Autonomous Systems [22] that clearly demands that autonomous systems must be made understandable during the engineering process. It must be noted that this type of transparency differs from what we consider in this contribution: We are interested in the resolution of conflict situations at run-time, through increasing transparency. Thus, a general system transparency contributes to situational transparency. There has been further discussion on "the right to explanation"[56] within the EU General Data Protection Regulation[4]. To that end, literature has identified self-explainability as a necessary non-functional system requirement within the last years [29, 8, 7]. Research on the validation of system explanations reveals that only explanations E which are provided in the right context C, to specific types of recipient groups G and with an appropriate means M are actually capable of increasing qualities like trust and safety [41, 13, 47]. Further on, diverse explainability requirements have been identified for diverse explainability stakeholder groups G [34].

We conclude that, for engineering explanations for risk mitigation in conflict situations, we must carefully take the needs of diverse conflict stakeholders, the explanation timing and the specific conflict context into account.

3 The Need for Explanations in Risk Mitigation

For the homologation of AVs, manufacturers are required to provide evidence that all risks are adequately mitigated, ensuring that the system is reasonably safe [14]. Conflict situations are often directly associated with increased risks, as the coordination with other traffic participants presents a major challenge for AVs as well as for human drivers and further vulnerable road users (VRUs).

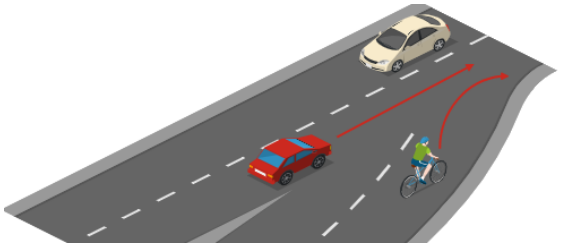
In order to understand the relevance of explanations as safety measures for diverse conflict situations, we examine three examples in the remainder of this section.

Example 1. Obstacle Ahead . Again consider the example from Fig. 1: We assume that the risk for all three behavioural options (a) evasive manoeuvre into oncoming traffic, (b) lane keeping with braking, and (c) evasive manoeuvre into the border (cf. Sect. 2.2) exceeds an acceptable level. Consequently, the system's homologation necessitates the implementation of safety measures to decrease the associated risk. Note that the manufacturer and, therefore, the AV, have limited control over factors such as roadside structure or animals appearing on the road: One solution could be to significantly restrict the operational design domain. However, this would strongly constrain the functionality of the system and thus it's real-world applicability. Thus, there is only a low potential to reduce the probability that such conflict situations occur. Moreover, the potential for reducing the severity of harm, e.g. by available technologies such as enhanced airbag systems, is often already

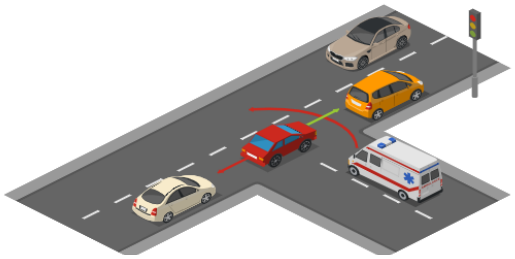
[4] `https://gdpr-info.eu/` (accessed on 19.12.2024)

exhausted. Efforts to improving the controllability of the situation, specifically for the behaviours (b) and (c), through adaptive braking behaviours or other internal means, may also be limited. It remains to consider safety measures for the behavioural option (a), where ego performs an evasive manoeuvre into the oncoming lane. The risk of this situation depends on the behaviour of the oncoming traffic. Therefore, resolving the conflict between ego and the oncoming traffic by means of explanations provides a viable possibility for risk mitigation.

Example 2. Lane Merging. A conflict situation that includes a vulnerable road user (VRU) is illustrated in Fig. 2a: A bicyclist drives on a ramp road and wants to enter the main road just in front of the ego vehicle. Ego has two options: (a) lane keeping and braking or (b) an evasive manoeuvre onto the oncoming traffic. We assume that both options are associated with high risks of injury due to collision of ego with the bicyclist or the car in oncoming traffic. Thus, safety measures must be implemented. The risk associated with this situation could again be mitigated by explanations. These explanations could either aim at preventing the bicyclist from entering the main road or again at warning other traffic participants of a potential evasive manoeuvre.



(a) Example 2 *lane merging*: a bicyclists wants to drive onto the main road from a ramp road.



(b) Example 3 *emergency vehicle*: AV A needs to either reverse or drive forward to let the emergency vehicle pass

Example 3. Emergency Vehicle. Fig. 2b presents another example; the ego vehicle is standing in front of a traffic light blocking the path of an emergency vehicle that attempts to enter the crossing to turn left. In this scenario, ego provokes the risk of fatal harm due to obstruction of an emergency vehicle. As there are vehicles in front and behind the ego vehicle, there is no possibility for ego to give way to the emergency vehicle. We again encounter a conflict situation, which in fact encompasses several individual conflicts between the different agents. The manufacturer of the ego vehicle mainly has two options to mitigate the risk associated with this situation: Either to prevent the occurrence of the situation or to provide the ego vehicle with means to encourage the vehicles in front or behind ego to make space via a targeted explanation.

Assuming that an integrated HARA conducted in compliance with ISO 26262 and ISO 21448 identifies various hazardous events involving conflicts such as those presented in the examples above. Then each of the hazardous events has to be further investigated. This includes analysing causes such as system failures or functional insufficiency as well as the evaluation of associated risks. If a comprehensive analysis of a hazardous conflict scenario yields that the associated risk is unreasonable, safety measures must be integrated into the system design and their efficiency must be proven.

Examples 1 to 3 demonstrate that solving conflicts may provide viable solutions to mitigate risks in such hazardous conflict scenarios. In the following sections, we examine in detail whether and

how explanations can be used as safety measures to reduce the entailed risk of hazardous conflict sufficiently.

4 Assessment of Risk Reduction through Explanations

Risk mitigation in conflict scenarios requires systematically integrating safety measures during the system design. We propose a structured design approach to assess early in the system development process whether explanations can be used as a safety measure to reduce the risk to a tolerable level. The approach is applied after a hazard and risk analysis (HARA) has been performed. The HARA provides us with a list of abstract scenarios that describe conflicts, in which our agent needs a strategy to coordinate with other agents to reduce the associated risks (cf. Sect. 3). The approach assesses whether explanations can be used as safety measures in the conflict scenarios to increase the likelihood of appropriate resolutions such that the risk is reduced to a tolerable level. The approach identifies key assumptions about the environment of *ego* and derives requirements for explanations that, when implemented, form the foundation for a safety case. With other words, the subsequent design process must ensure that *ego* will give an appropriate explanation. We demonstrate this approach using Example 1, Obstacle Ahead from Sect. 1.

Since we frame conflicts as situations in a game (cf. Sect. 2.1), we also analyse the effect of explanations in conflict scenarios as games. In a nutshell, the game captures the two conflicting parties with their goals and beliefs and how they choose their actions. The question "Can an (idealised) explanation sufficiently reduce the risk?" is answered with "yes" if modifying the game by introducing an explanation, leads to the desired resolution of the conflict.

We define guiding questions to frame the conflict situation as a game. We determine whether an explanation changes the game in such a way that the likelihood of a conflict resolution increases and thus the risk is sufficiently reduced. Our approach is inspired by the PARTS method from Brandenburger [5]. PARTS was originally developed to improve a company's competitive situation. The current situation is analysed as a game and it is investigated how the game can be changed so that the company ends up in a more competitive situation.

Fig. 3 provides an overview of our approach. We assume a set of potentially hazardous conflict scenarios to be given and examine the scenarios one by one. In Step 1, the engineers identify potential resolutions of the conflict, focusing on cases where explanations could sufficiently reduce the risk. In Step 2, the potential other players are characterised. In Step 3, games for the different player types are analysed to determine in which games *ego*'s desired course of action leads to a hazard. In Step 4, the game is modified; *ego* can give idealised explanations. It is examined whether the risk is thereby sufficiently reduced.

4.1 Step 1: Identify Resolutions

The engineers engage in brainstorming to explore resolutions for the conflict scenario. Note that this activity might even expand the design space since it is performed at an early stage of the development process. The engineers in particular assess whether resolutions of the conflict might be achieved through explanations. The resolutions that might be achievable through explanations and that could reduce the risk to a tolerable level are collected in terms of (i) the preferred course of action of *ego* (course(*ego*)), (ii) the desired course of action of the recipient (course(*other*)), and (iii) the *confidence* (i.e. the required confidence that (ii) be realised). The engineers define a level of confidence for the resolution, ensuring it supports a safety case demonstrating the achievement of the required risk reduction.

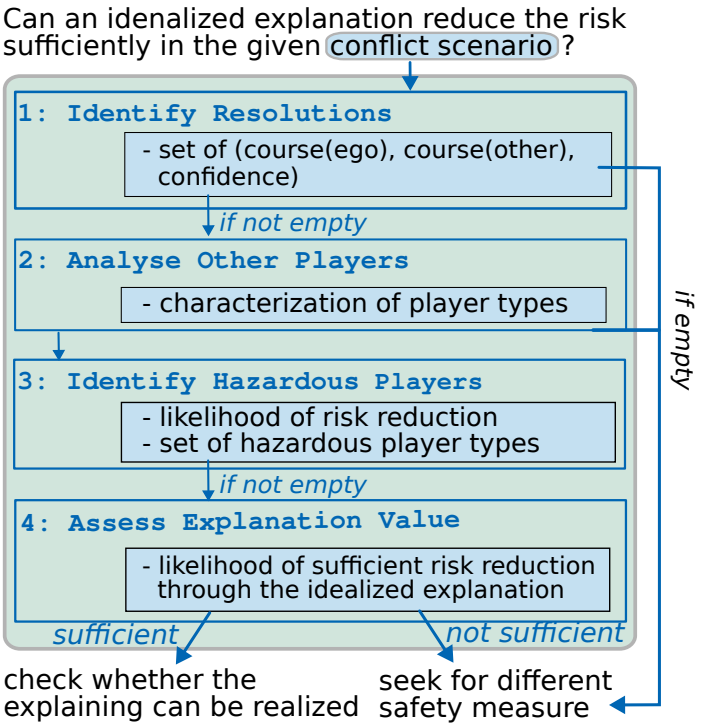


Fig. 3: Overview over our explanation value assessment approach.

Example 4. (Resolutions) *In our running example, we specify for instance*
*course(*ego*):=“A wants to drive onto the other lane and without collision”,*
*course(*other*):=“On-coming traffic drive at the right side of their lane”,*
confidence:=“The confidence in the effect has to be at least X.”

At the end of step `(Identify Resolutions)` we derived a set of resolutions where explanations could achieve a risk reduction to a tolerable level. If this set is not empty, we proceed to further analyse the scenario.

4.2 Step 2: Analyse Other Players

So far, the conflict scenario has been dealt with at a high abstraction level. Hence, player B usually represents entities from different behavioural classes. In this step, we define a set of games for the given conflict scenario by specifying the player type. A run corresponding to the hazardous conflict scenario is possible in each derived game. Later steps will examine whether explanations can rule out the conflicts and reduce the risk to a tolerable level.

To specify the games, we determine the actions, goals, beliefs and the level of cooperation/competition of each player type of B. Guiding questions are summarised in Fig. 4. Note that in our setting rationality is subjective. Player B is called rational if it selects the course of action that maximises its expected payoff – logically consistent with all its available information. Degrees of rationality emerge when the decision is not logically consistent.

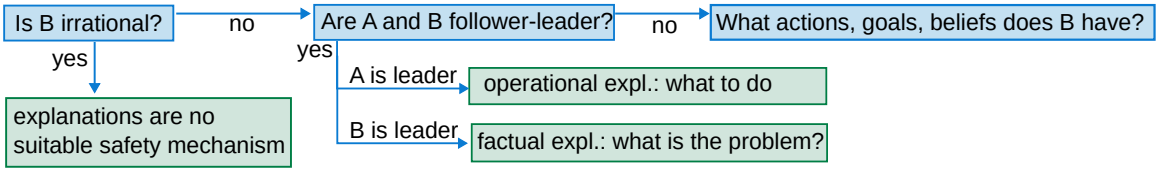


Fig. 4: What kind of player is B? What are B's goals?

While our approach is applied at design time, it must be acknowledged that during run-time, i.e. when the explanation will be given, only “What does the ATA A know about the actions, goals and beliefs of B?” can be answered. To construct a safety case, we must consider the uncertainty of information when answering the questions of Fig. 4. For reasons of brevity, we do not discuss this issue further and only mention here that we envision modelling the ground truth and A's perspective via an epistemic game (i.e. we explicitly model A's and B's beliefs).

Example 5. (Player Analysis) *In our running example, we must consider the possible types of B. For our example, we assume that B can be (i) an ATA of the same manufacturer (ii) an ATA certified to implement an ATACode, (iii) an uncertified ATA or (iv) a vehicle driven by a human. An ATA implementing the ATACode is guaranteed to have the top priority goals g_{lifes} :=“saving lives” and g_{help} :=“being helpful, if possible”. We consider none of the player types as irrational (but limited rational) and we are not in a follower-leader relation with any of them.*

Regarding their goals, in case (i-ii), we derive that B has the top priority goals g_{lifes} and g_{help} In cases (iii-iv), we cannot take these goal priorities for granted. For (iv), i.e. cars with human drivers, we assume that their goals are g_{lifes}, g_{help}, g_{col}:=“avoid collisions”, g_{dam} :=“avoid damages” –

in decreasing order of priority. Regarding B's action, we assume standard vehicle dynamics and car-to-car communication between A and B in cases (i-iii). In case (i), we moreover know the internal representation of the B's situational assessment.

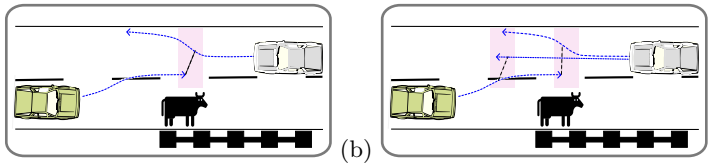
At the end of step `(Analyse Other Players)`, we have specified what different types of players A may encounter in the conflict scenario. In particular, it is captured what A thinks the situation is like including what B's goals are, what B's priorities of goals are, what the actions of B are and how B chooses its strategy (i.e. its type of rationality).

4.3 Identify Hazardous Players

In this step, it is determined which player type poses a risk, given all possible resolutions of step `(Identify Resolutions)`. If no resolution leads to a sufficient risk reduction, the next step `(Assess Explanation Value)` assesses whether an explanation can do so.

For a given resolution R, all the games are examined where A implements R and interacts with a possible player B of step `(Analyse Other Players)`. Based on B 's player type we determine B's potential strategies and assign likelihoods to them. Then the risk of the resolution is assessed. We envision that the risk is judged based on the ground truth effect that is captured as part of an epistemic game (cf. `(Analyse Other Players)`). To simplify the discussion, we assume that it is ensured by design that the ground truth and A's beliefs are well-aligned so that a distinction between ground truth and A's beliefs is not necessary.

Example 6. (Hazardous Player Types) Fig. 5 illustrates that the player type influences what is considered as likely: The certified ATA will likely try to give way to A, but may not be able to react in time, while the uncertified ATA might even be speeding and ignore that A is approaching.



(a) (b)

Fig. 5: If the AV decides to avoid the obstacle by entering the neighbouring lane, the likely reaction of a certified ATA (a) and an uncertified ATA (b) is different.

At the end of step `(Identify Hazardous Players)` we have specified assumptions about the likelihood of each possible player type of B choosing certain strategies. Additionally, we have identified the player types, for which the resulting risk is not tolerable under any resolution.

4.4 Assess Explanation Value

In this step, we examine what the added value of an explanation is in terms of risk reduction. The goal of this step is to answer whether an honest explanation can in principle reduce the risk by increasing the probability of at least one of the desired resolutions. Therefore the effect of idealised explanations in the conflict scenario is examined. We make the following idealising assumptions that

(i) A is aware of its beliefs, goals and the intended resolution

(ii) A can frame this information via an explanation such that B can integrate it into its world model (without conflicts) and
(iii) this explaining is instantaneous.

In the subsequent system development process, the assumptions (i-iii) must be replaced by realistic requirements, if possible. Otherwise explaining is considered not to be feasible.

For all hazardous player types of B, respective games of A and B are examined, but the game is modified by letting A have the action of idealised explaining. It is assessed whether the risk associated with one of A's resolutions becomes tolerable by providing an idealised explanation.

Example 7. (Added Value) In case (a) of Example 6, an idealised explanation would cause B to give way since its goal is being helpful. In case (b) though, an explanation might not cause B to give way. Since B is not certified, it is less clear what its goals are, hence e.g. might accept severe damages as long as its own functionality is not compromised.

At the end of step `(Assess Explanation Value)` we have an assessment of how likely the idealised explanation reduces the risk to a tolerable level for each resolution. If there is no resolution with a tolerable risk, other safety measures must be identified. Otherwise, the system design continues. It must ensure that *ego* will give an appropriate explanation.

4.5 Discussion & Conclusion

The presented steps target the analysis of the other player B, i.e. conflict party, that is player B. Analysing whether an idealised explanation can reduce the risk of the conflict sufficiently, is just one step towards actually realising such a system. In this section, we considered a simplified setting as we are targeting the early system development phase. We hence did not discuss issues e.g. caused by wrong beliefs or scenarios where A and B can resolve a conflict by exchanging information.

5 Engineering Run-time Explainability for Risk Mitigation

We have identified the potential and need of explanations in risk mitigation in the previous sections. To actually integrate explanations into conflict situations, we must discuss means for explaining *at run-time*, i.e., explaining in conflict situations. The difficulty of run-time explainability, especially in conflict situations, is the dynamic nature of run-time reasoning: the situational context is switching constantly, and decisions must be made within very short time frames. Thus, a self-explainable system must be enabled to deliver explanations fast at run-time. Specifically for taming the dynamic nature of run-time explainability, the MAB-EX framework [4] has been introduced. MAB-EX assumes the existence of explanation models which we motivate within our application case of risk mitigation in Sect. 5.1, before we give details for the MAB-EX phases in Sect. 5.2. We infuse MAB-EX with our requirements for explanations for risk mitigation in Sect. 5.3.

5.1 Explanation Models: Enabling Timely Explanations during Run-time

Explanation models have been introduced before [18] to speed up run-time explanation processes: Pre-built at design time from existing system models, explanations can be derived quickly from such explanation models during run-time. Exemplary source system models can include diagrams and program code. A process for deriving explanation models from timed automata has been discussed in [50]. Through the formal nature of the system models that it is extracted from, an explanation model is formalised and machine-readable itself. Thus, an explanation model can be considered an

intermediate explanation format where internal explanations can be retrieved from. These internal explanations are not yet in a format that is understandable by respective recipient stakeholders G.

Consider again the obstacle evasion Example 1 from Fig. 1. We depict an exemplary cut-out of an explanation model for the action *evade into the oncoming traffic* in Fig. 6. The *explanation path* that is highlighted could represent the logical formula

$$because(\mathsf{evade}_{\mathsf{ot}}, and(\mathsf{p}_4 \wedge \neg \mathsf{p}_5 \wedge \neg \mathsf{p}_6)).$$

In this example the symbols p_4, p_5 and p_6 abstractly formalise properties of the conflict. p_4 could mean that oncoming lane is relatively free, $\neg p_5$ could describe that the obstacle is not within a safe braking distance and $\neg p_6$ could describe that the road side is not free.

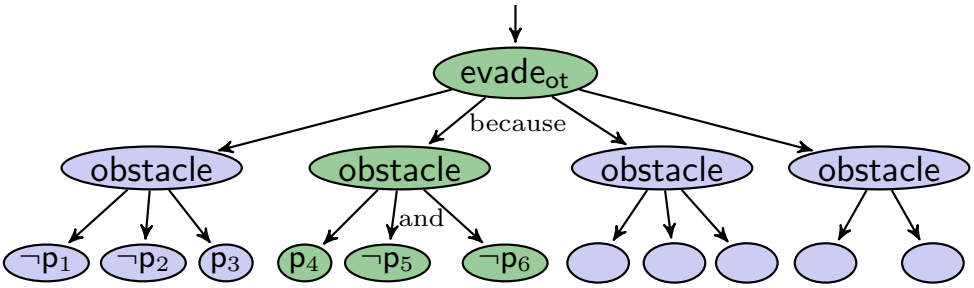


Fig. 6: Explanation model EM for the action "evade into the oncoming traffic" for Example 1 (cf. Fig. 1).

This internal, logical, explanation must be translated to a format that is understandable by the targeted recipient group G. For a passenger of the ego AV, the internal explanation could be translated into a counterfactual explanation in natural language: *The AV is evading into the oncoming traffic, as an obstacle is ahead, and the options of braking and driving into the obstacle or evading into the lane border are not suitable.* Which of the potential explanation paths in EM must be retrieved, depends on the specific conflict context C (i.e. the specific traffic situation). In another conflict situation, it might be that the AV evades into the oncoming traffic for different reasons. Thus, a different explanation path must be considered.

As explanation models are retrieved from arbitrary large and many system models, they can be overly complex, containing information that is not necessary for their specific explanandum X, explanation context C and the recipient stakeholders G. Thus, [50] suggests measures for simplification of explanation models. The intuition is that, e.g., an end-user might need to know less detailed information then an engineer who intends to debug the system. In the case of explanations for risk mitigation, we can explore what can be omitted in the explanation model based on the idealised explanation that has been identified before in Sect. 4.

The key benefit of the explanation model is that it is generated at design time, thus enabling the derivation of explanations at run-time quickly. Of course, this means that the explanation model must be updated at run-time, e.g. when a system update is necessary. This issue has been addressed in [50] and we omit further details on it for brevity. We now step from the design-time explanation model construction to run-time explainability: The process of explaining. I.e., which explanation path must be extracted when, and how it should be translated. This explanation process is captured by the MAB-EX framework.

5.2 MAB-EX Framework: Taming Run-time Explainability

MAB-EX is a modular framework that allows to integrate self-explainability capabilities into already existing systems. It supports the system in the entire process of self-explanation at run-time. MAB-EX is inspired by the MAPE-K Loop for self-adaptive systems [53] and follows a cycle of four phases at run-time (cf. Fig. 7): In the **M**onitoring phase, the system and its environment is observed, including all involved stakeholders. Next, in the **A**nalysis phase, the data from the previous phase is examined for unexpected system behaviour or other reasons why an explanation might be necessary. Within the **B**uild phase, an internal, machine-readable, explanation is derived from pre-built explanation models. Finally, in the phase **EX**plain, the internal explanation is translated w.r.t. explanation recipients G.

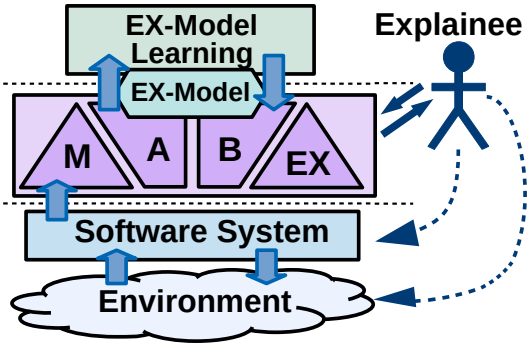


Fig. 7: MAB-EX Framework from [4].

In the Example 1, the ego AV must **M**onitor the explanation context C and all involved stakeholders; These could include passengers or drivers of the AV, as well as the agent on the lane for oncoming traffic. A result of the **A**nalysis phase could be that ego can conclude that the other agent does not have sufficient knowledge about the conflict and ego's options and must be informed. An explanation is extracted from explanation models in the **B**uild phase (cf. Sect. 5.1). Depending on the recipient stakeholder group G, the explanation must be translated with different explanation means M in the **EX**plain phase.

5.3 Challenges in Refining MAB-EX towards Explanations for Risk Mitigation

The phases of MAB-EX provide a framework skeleton for engineering run-time explainability. These phases must be infused with actionable requirements to pave the way towards integrating explanations into risk mitigation approaches. For this, we discuss key questions for each phase of MAB-EX using the characteristics of risk-mitigating explanations that have been identified in Sect. 4. This subsection can be considered as an outlook towards engineering run-time explanations for risk mitigation.

Monitoring. *What information does ego need to monitor to assess whether an explanation can be helpful for risk mitigation?*

The key assumptions about ego's environment must be captured, so that those types of explanations that have been identified as helpful for risk mitigation in the previous section can be provided. Our explanation context C is a conflict situation. In our case of conflicts that are modelled as games,

what needs to be monitored includes the beliefs of ego, the level of certainty of those beliefs and the knowledge of the nature of other players (e.g. cooperative vs competitive). Note that we monitor information about the explanation context (i.e. the conflict), without assessing the need for an explanation in this phase.

Analysis. *Can an explanation help in reducing risk in the conflict situation? Which agents are involved and need an explanation?*

In step one `(Identify Resolutions)` of our design approach, specific conflict resolutions in which explanations can be beneficial have been identified. This knowledge is used in the Analysis phase of MAB-EX, to identify the need of an explanation for risk mitigation within the monitored data. As the process of identifying such resolutions through the hazard and risk analysis (HARA) is done at design time, the analysis phase can be quickly done at run-time.

Build. *Can ego provide the necessary information? What does an explanation recipient B need to know when?*

In the case of risk mitigation a prerequisite for our explanation model from Sect. 5.1 is that it can provide the explanation content that has been identified as necessary in Sect. 4. This means that the process of deriving and refining an optimal explanation model is a process that runs in parallel to the design approach that was sketched in Sect. 4.

EXplain. *How much time do we have for the explanation? What type of explanation is suitable? What is the optimal explanation means?*

Based on the interface specification of a certified B we can determine an explanation encoding and derive time windows that B guarantees for incorporating the content into its world model. In general, but especially also for explaining for risk mitigation, this phase depends highly on the type of recipients G. We consider both autonomous and human agents in this paper. Especially in the case of human agents, real guarantees for successfully resolving the conflict through a delivered explanation are difficult to provide. We shed light on some specific challenges that come with explaining to humans in the next section, also deriving some explanation requirements that must be met to ensure that explanations, also for humans, can actually be helpful for risk mitigation.

6 Challenges and Requirements in Explaining to Humans

Explanations act as a bridge between the system's decision-making process and the human agent. Thus, they are central to gaining the trust of humans into a system, particularly an autonomous one. They serve as a mechanism for transferring knowledge, clarifying decisions, and justifying actions. In Sect. 4, we explored the various challenges and requirements for providing explanations to ATAs. However, as mentioned in Sect. 5, the recipients of these explanations could also be human agents (HAs). We elaborate on challenges that come specifically with explaining to humans in the following Sect. 6.1 and derive some requirements from this in Sect. 6.2.

6.1 Challenges

With explanations to humans for conflict resolution, it is essential that the explanation accounts for human *cognitive limitations*, *contextual variability*, and the *system opaqueness and complexity* of the system providing the explanation.

Cognitive limitations ($Ch1$). Humans have limited cognitive resources, including memory, attention span, and information processing capacity, which can hinder their ability to understand complex explanations [54]. Additionally, humans tend to selectively focus on information that they deem most relevant; For humans to be able to grasp the explanation provided, the explanation must

align closely with the recipient's immediate goals and priorities [28]. Consider the lane merging example in Sect. 3, where an explanation such as *"Please do not merge as this car cannot stop in time"* is more cognitively demanding for the bicyclist than a simpler statement like *"Don't merge!"*.

Contextual variability ($Ch2$). Dynamic evolution of the explanation context also add to the challenges of explaining. In safety-critical situations, like an imminent emergency manoeuvre, explanations must be concise, actionable, and on-point. Considering the Obstacle Ahead example from Sect. 1, a concise explanation like *"Evading the obstacle in front"* may suffice for passengers at the moment. At the same time, a more detailed post hoc report may be necessary for regulatory purposes (cf. [46]). Apart from the context, explanation timing plays a major role, as we also observed in the previous section. A timely explanation not only increases the trust and preference for the system [30, 13, 49], but also helps in a better understanding of the situation [32]. Previous works of Bairy and Fränzle show us how the optimal time for providing an explanation can be determined based on the attention level of the human [1, 2].

System opaqueness and complexity ($Ch3$). Autonomous traffic agents (ATAs) are generally complex systems and often not completely transparent by design. Many ATAs also include self-learning AI components, that are highly opaque and known for their lack of interpretability [36]. This limits their potential to explain decisions in a comprehensible way. The authors of [39] argue that effective explanation methods for AI systems must account for user context, mental models, and the social dynamics of trust, moving beyond merely algorithmic transparency. Additionally, explanation scalability and personalisation pose a huge challenge; systems must provide explanations to diverse stakeholders with varying levels of knowledge and preferences. There are a few models which tackle this exact problem [50, 44].

6.2 Requirements

The challenges ($Ch1 - Ch3$) reveal obstacles that need to be addressed in order to provide an explanation that is valuable to humans. Addressing these challenges effectively requires a strategic approach, and the requirements ($R1 - R3$) outlined below aim to create a structured pathway to overcome these issues.

Clarity and simplicity ($R1$). The first and foremost requirement is *clarity and simplicity*, which deals with $Ch1$. Explanations should avoid unnecessary jargon and be presented in a structured manner that facilitates understanding. In practice, this might involve using plain language explanations for passengers, e.g., "*Braking to maintain a safe distance*"; and "*Don't merge!*" for the bicyclist. The explanation should also consider the *emotional/cognitive state* of the human. Under normal conditions, the human working memory can hold only about 7 (± 2) items at a time [38]. Emotions, stress, and a lot of other factors can negatively affect the working memory, leading to a lower retention rate (than the normal 7 (± 2) items) of the information in the human [3].

Relevance and Timing ($R2$). Another critical requirement is *relevance*. An effective explanation is one which aligns with the specific needs and goals of the human. *Timeliness* is another critical requirement, particularly in time-sensitive scenarios. Relevance and timing go hand-in-hand. In safety-critical scenarios, such as in the Obstacle Ahead Example from section 1, an explanation, such as "*Obstacle detected ahead*", should be provided quickly enough to support real-time decision-making. Relevance and timing of an explanation are the requirements for $Ch2$, which deals with the contextual variability of an explanation.

Stakeholder needs ($R3$). Stakeholder needs build a core requirement to deal with Challenge $Ch3$. Explanations must be able to adapt to different stakeholders' needs. Also, an explanation provided, in case of an accident, to a passenger might not need all the information that the developers/law enforcement require. For example, consider the Lane Merging Example from section 3, a passenger's

explanation might prioritise safety reasoning, the bicyclist explanation would centre on expressing (indirectly) AV's intended next move, whereas a developer's explanation focuses on debugging information, and an explanation to law enforcement would focus on all the above-mentioned details along the information about the other vehicles involved in the accident.

In summary, providing explanations to humans presents unique challenges compared to explanations for other ATAs. By addressing these challenges through well-defined requirements, AVs can ensure that their explanations not only help maintain —or even enhance— the trust of the human agents in the system but also foster a deeper understanding of the system's decisions and behaviours.

7 Conclusion

We contribute to the field of risk mitigation for AVs by advocating the need for explanations in high-risk conflict situations. Trough state of the art approaches within our three key topics *conflict*, *risk* and *explanation*, we substantiate our findings. By methodically analysing risk scenarios in a game-theoretic setting, we emphasise the value that an explanation can have in risk mitigation. For this, we take different types of players into account and argue that explanations for specific player types can indeed reasonably reduce risks. We discuss how to introduce risk mitigation through explanations into frameworks for run-time explainability, thereby paving the way towards implementing our approach. We discuss specific research questions and challenges for engineering run-time explainability for risk-mitigation and focus specifically on the complex challenge of explaining to humans.

While the contribution at hand advocates the crucial role of explanations in risk mitigation, next steps would be to implement our findings and to analyse them in practice. For this, a user study within a simulation environment would be reasonable, e.g. starting from the example scenarios that we have identified in Sect. 3.

Acknowledgments. This research was supported by the Innovation Campus for Future Mobility (`www.icm-bw.de`) and by the Helmholtz Association within the Core Informatics project.

References

[1] Akhila Bairy and Martin Fränzle. "Efficiently Explained: Leveraging the SEEV Cognitive Model for Optimal Explanation Delivery". In: *Applied Human Factors and Ergonomics (AHFE 2024)* 148 (2024). DOI: 10.54941/ahfe1005221.

[2] Akhila Bairy and Martin Fränzle. "Optimal Explanation Generation Using Attention Distribution Model". In: *Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications* 70.70 (2023). DOI: 10.54941/ahfe1002928.

[3] Rachael N. Blasiman and Christopher A. Was. "Why Is Working Memory Performance Unstable? A Review of 21 Factors". In: *Europe's Journal of Psychology* 14.1 (Mar. 2018), pp. 188–231. DOI: 10.5964/ejop.v14i1.1472. URL: https://ejop.psychopen.eu/index.php/ejop/article/view/1472.

[4] Mathias Blumreiter, Joel Greenyer, Francisco Javier Chiyah Garcia, Verena Klös, Maike Schwammberger, Christoph Sommer, Andreas Vogelsang, and Andreas Wortmann. "Towards Self-Explainable Cyber-Physical Systems". In: *22nd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion*. 2019, pp. 543–548. DOI: 10.1109/MODELS-C.2019.00084.

[5] Adam M. Brandenburger and Barry J. Nalebuff. "The Right Game: Use Game Theory to Shape Strategy. (Cover story)". In: *Harvard Business Review* 73.4 (1995), pp. 57–71. ISSN: 00178012.

[6] Giuseppe Cartella, Marcella Cornia, Vittorio Cuculo, Alessandro D'Amelio, Dario Zanca, Giuseppe Boccignone, and Rita Cucchiara. "Trends, Applications, and Challenges in Human Attention Modelling". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. Ed. by Kate Larson. Survey Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, pp. 7971–7979. DOI: 10.24963/ijcai.2024/882. URL: https://doi.org/10.24963/ijcai.2024/882.

[7] Larissa Chazette. "Requirements engineering for explainable systems". PhD thesis. University of Hanover, Hannover, Germany, 2023. URL: https://www.repo.uni-hannover.de/handle/123456789/13370.

[8] Larissa Chazette, Wasja Brunotte, and Timo Speith. "Explainable software systems: from requirements analysis to system evaluation". In: *Requir. Eng.* 27.4 (2022), pp. 457–487. DOI: 10.1007/s00766-022-00393-5. URL: https://doi.org/10.1007/s00766-022-00393-5.

[9] Larissa Chazette, Wasja Brunotte, and Timo Speith. " Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue ". In: *2021 IEEE 29th International Requirements Engineering Conference (RE)*. Los Alamitos, CA, USA: IEEE Computer Society, Sept. 2021, pp. 197–208. DOI: 10.1109/RE51729.2021.00025. URL: https://doi.ieeecomputersociety.org/10.1109/RE51729.2021.00025.

[10] Hoong-Chor Chin and Ser-Tong Quek. "Measurement of traffic conflicts". In: *Safety Science* 26.3 (1997), pp. 169–185. ISSN: 0925-7535. DOI: 10.1016/S0925-7535(97)00041-6.

[11] Werner Damm, Martin Fränzle, Willem Hagemann, Paul Kröger, and Astrid Rakow. "Dynamic Conflict Resolution Using Justification Based Reasoning". In: *Proceedings of the 4th Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology, CREST@ETAPS 2019, Prague, Czech Republic, 7th April 2019*. Ed. by Georgiana Caltais and Jean Krivine. Vol. 308. EPTCS. 2019, pp. 47–65. DOI: 10.4204/EPTCS.308.4. URL: https://doi.org/10.4204/EPTCS.308.4.

[12] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. "Multi-Agent Systems: A Survey". In: *IEEE Access* 6 (2018), pp. 28573–28593. DOI: 10.1109/ACCESS.2018.2831228.

[13] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert. "Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload". In: *Transportation Research Part C: Emerging Technologies* 104 (2019), pp. 428–442. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.05.025. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18313640.

[14] European Union. *Commission Implementing Regualtion (EU)2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles*. Legal Act. 2022.

[15] Andrea Ferrario and Michele Loi. "How Explainability Contributes to Trust in AI". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1457–1466. ISBN: 9781450393522. DOI: 10.1145/3531146.3533202. URL: https://doi.org/10.1145/3531146.3533202.

[16] Mahdi Gabaire, Haniyeh Ghomi, and Mohamed Hussein. "Investigating the contributing factors to autonomous Vehicle-Road user Conflicts: A Data-Driven approach". In: *Accident Analysis & Prevention* 211 (2025), p. 107898. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2024.107898. URL: https://www.sciencedirect.com/science/article/pii/S0001457524004433.

[17] Johan Galtung. "Violence, Peace, and Peace Research". In: *Journal of Peace Research* 6.3 (1969), pp. 167–191. DOI: 10.1177/002234336900600301.

[18] Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patrón, and Helen F. Hastie. "Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots". In: *CoRR* abs/1803.02088 (2018). arXiv: 1803.02088. URL: http://arxiv.org/abs/1803.02088.

[19] Jeffrey K. Gurney. "Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles". In: 2013. URL: https://api.semanticscholar.org/CorpusID:107966361.

[20] Md Mohasin Howlader, Fred Mannering, and Md Mazharul Haque. "Estimating crash risk and injury severity considering multiple traffic conflict and crash types: A bivariate extreme value approach". In: *Analytic Methods in Accident Research* 42 (2024), p. 100331. ISSN: 2213-6657. DOI: https://doi.org/10.1016/j.amar.2024.100331. URL: https://www.sciencedirect.com/science/article/pii/S2213665724000150.

[21] Yong K. Hwang and Narendra Ahuja. "Gross motion planning – a survey". In: *ACM Comput. Surv.* 24.3 (1992), pp. 219–291. ISSN: 0360-0300. DOI: 10.1145/136035.136037.

[22] "IEEE Standard for Transparency of Autonomous Systems". In: *IEEE Std 7001-2021* (2022), pp. 1–54. DOI: 10.1109/IEEESTD.2022.9726144.

[23] International Organization for Standardization. *ISO 21448: Road vehicles – Safety of the intended functionality*. Standard. 2022.

[24] International Organization for Standardization. *ISO 26262: Road vehicles – Functional safety*. Standard. 2018.

[25] International Organization for Standardization. *ISO/IEC Guide 51: Safety aspects — Guidelines for their inclusion in standards*. Standard. 2014.

[26] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. "The social dilemma of autonomous vehicles". In: *Science* 352.6293 (2016), pp. 1573–1576. ISSN: 0036-8075. DOI: 10.1126/science.aaf2654. eprint: http://science.sciencemag.org/content/352/6293/1573.full.pdf. URL: http://science.sciencemag.org/content/352/6293/1573.

[27] John D. Lee and Katrina A. See. "Trust in Automation: Designing for Appropriate Reliance". In: *Human Factors* 46.1 (2004), pp. 50–80. DOI: 10.1518/hfes.46.1.50{\textunderscore}30392.

[28] Daniel Kahneman. *Attention and Effort*. Prentice-Hall, 1973.

[29] Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. "Explainability as a Non-Functional Requirement". In: *27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019*. Ed. by Daniela E. Damian, Anna Perini, and Seok-Won Lee. IEEE, 2019, pp. 363–368. DOI: 10.1109/RE.2019.00046. URL: https://doi.org/10.1109/RE.2019.00046.

[30] Jeamin Koo, Dongjun Shin, Martin Steinert, and Larry Leifer. "Understanding driver responses to voice alerts of autonomous car operations". In: *International Journal of Vehicle Design* 70 (Jan. 2016), p. 377. DOI: 10.1504/IJVD.2016.076740.

[31] Philip Koopman and Michael Wagner. "Autonomous Vehicle Safety: An Interdisciplinary Challenge". In: *IEEE Intelligent Transportation Systems Magazine* 9.1 (2017), pp. 90–96. DOI: 10.1109/MITS.2016.2583491.

[32] Moritz Körber, Lorenz Prasch, and Klaus Bengler. "Why Do I Have to Drive Now? Post Hoc Explanations of Takeover Requests". In: *Hum. Factors* 60.3 (2018), pp. 305–323. DOI: 10.1177/0018720817747730.

[33] Birte Kramer, Christian Neurohr, Matthias Büker, Eckard Böde, Martin Fränzle, and Werner Damm. "Identification and Quantification of Hazardous Scenarios for Automated Driving". In: *Model-Based Safety and Assessment - 7th International Symposium, IMBSA 2020, Lisbon,*

Portugal, September 14-16, 2020, Proceedings. Ed. by Marc Zeller and Kai Höfig. Vol. 12297. Lecture Notes in Computer Science. Springer, 2020, pp. 163–178. DOI: 10.1007/978-3-030-58920-2_11. URL: https://doi.org/10.1007/978-3-030-58920-2%5C_11.

[34] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research". In: *Artificial Intelligence* 296 (2021), p. 103473. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103473. URL: https://www.sciencedirect.com/science/article/pii/S0004370221000242.

[35] Patrick Lin. "Why Ethics Matters for Autonomous Cars". In: *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Ed. by Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 69–85. ISBN: 978-3-662-45854-9. DOI: 10.1007/978-3-662-45854-9_4. URL: https://doi.org/10.1007/978-3-662-45854-9_4.

[36] Zachary C. Lipton. "The mythos of model interpretability". In: *Commun. ACM* 61.10 (Sept. 2018), pp. 36–43. ISSN: 0001-0782. DOI: 10.1145/3233231. URL: https://doi.org/10.1145/3233231.

[37] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies". In: *Journal of Biomedical Informatics* 113 (2021), p. 103655. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2020.103655. URL: https://www.sciencedirect.com/science/article/pii/S1532046420302835.

[38] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956), p. 81. DOI: 10.1037/h0043158.

[39] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences". In: *CoRR* abs/1712.00547 (2017). arXiv: 1712.00547. URL: http://arxiv.org/abs/1712.00547.

[40] Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991. ISBN: 9780674341166. URL: http://www.jstor.org/stable/j.ctvjsf522.

[41] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI". In: *ACM Comput. Surv.* 55.13s (July 2023). ISSN: 0360-0300. DOI: 10.1145/3583558. URL: https://doi.org/10.1145/3583558.

[42] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior (60th-Anniversary Edition)*. Princeton University Press, 2007. ISBN: 978-0-691-13061-3. URL: http://www.jstor.org/stable/j.ctt1r2gkx.

[43] S. R. Perkins and J. I. Harris. "Traffic Conflict Characteristics – Accident Potential at Intersections". In: *Highway Research Record* (225 1967), pp. 35–43.

[44] Alun D. Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. "Stakeholders in Explainable AI". In: *CoRR* abs/1810.00184 (2018). arXiv: 1810.00184. URL: http://arxiv.org/abs/1810.00184.

[45] P. Raja and S. Pugazhenthi. "Optimal path planning of mobile robots: A review". In: *International Journal of the Physical Sciences* 7 (9 Feb. 2012), pp. 1314–1320. DOI: 10.5897/IJPS11.1745.

[46] Astrid Rakow, Mehrnoush Hajnorouzi, and Akhila Bairy. "What to tell when? - Information Provision as a Game". In: *Proceedings Fifth International Workshop on Formal Methods for Autonomous Systems, FMAS@iFM 2023, Leiden, The Netherlands, 15th and 16th of November*

2023. Ed. by Marie Farrell, Matt Luckcuck, Mario Gleirscher, and Maike Schwammberger. Vol. 395. EPTCS. 2023, pp. 1–9. DOI: 10.4204/EPTCS.395.1. URL: https://doi.org/10.4204/EPTCS.395.1.

[47] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi. ACM, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[48] Jackeline Rios-Torres and Andreas A. Malikopoulos. "A Survey on the Coordination of Connected and Automated Vehicles at Intersections and Merging at Highway On-Ramps". In: *IEEE Transactions on Intelligent Transportation Systems* 18.5 (2017), pp. 1066–1077. DOI: 10.1109/TITS.2016.2600504.

[49] Peter A. M. Ruijten, Jacques M. B. Terken, and Sanjeev Chandramouli. "Enhancing Trust in Autonomous Vehicles through Intelligent User Interfaces That Mimic Human Behavior". In: *Multimodal Technol. Interact.* 2.4 (2018), p. 62. DOI: 10.3390/mti2040062.

[50] Maike Schwammberger and Verena Klös. "From Specification Models to Explanation Models: An Extraction and Refinement Process for Timed Automata". In: *Proceedings Fourth International Workshop on Formal Methods for Autonomous Systems (FMAS) and Fourth International Workshop on Automated and verifiable Software sYstem DEvelopment (ASYDE), FMAS/ASYDE@SEFM 2022, and Fourth International Workshop on Automated and verifiable Software sYstem DEvelopment (ASYDE)Berlin, Germany, 26th and 27th of September 2022*. Ed. by Matt Luckcuck and Marie Farrell. Vol. 371. EPTCS. 2022, pp. 20–37. DOI: 10.4204/EPTCS.371.2. URL: https://doi.org/10.4204/EPTCS.371.2.

[51] Maike Schwammberger, Raffaela Mirandola, and Nils Wenninghoff. *Explainability Engineering Challenges: Connecting Explainability Levels to Run-Time Explainability*. Ed. by Luca Longo, Sebastian Lapuschkin, and Christin Seifert. 2024. DOI: 10.1007/978-3-031-63803-9_11. URL: https://doi.org/10.1007/978-3-031-63803-9%5C_11.

[52] Guni Sharon, Roni Stern, Ariel Felner, and Nathan Sturtevant. "Conflict-Based Search For Optimal Multi-Agent Path Finding". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 26.1 (2012), pp. 563–569. ISSN: 2159-5399. DOI: 10.1609/aaai.v26i1.8140.

[53] David Sinreich. "An architectural blueprint for autonomic computing". In: *IBM Autonomic Computing – White Paper* (2006).

[54] John Sweller. "Cognitive load during problem solving: Effects on learning". In: *Cognitive Science* 12.2 (1988), pp. 257–285. ISSN: 0364-0213. DOI: https://doi.org/10.1016/0364-0213(88)90023-7. URL: https://www.sciencedirect.com/science/article/pii/0364021388900237.

[55] Underwriter Laboratories. *ANSI/UL 4600 - Standard for Evaluation of Autonomous Products*. Standard. 2022.

[56] Sandra Wachter, Brent Mittelstadt, and Chris Russell. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. arXiv: 1711.00399 [cs.AI]. URL: https://arxiv.org/abs/1711.00399.

[57] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. "From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI". In: *CoRR* abs/2105.05424 (2021). arXiv: 2105.05424. URL: https://arxiv.org/abs/2105.05424.