Carl von Ossietzky Universität Oldenburg

Fakultät II - Informatik, Wirtschafts- und Rechtswissenschaften
Department für Informatik

Optimising Timing of Explanations in Autonomous Vehicles

Von der Fakultät für Informatik, Wirtschafts- und Rechtswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

angenommene Dissertation
von: Akhila Bairy
geboren am 10.05.1994 in Bengaluru, Indien

Erstgutachter:	Prof. Dr. Martin Georg Fränzle
Zweitgutachter:	Prof. Dr. Christoph Herrmann
Datum der Einreichung:	28. Feb 2025
Datum der Verteidigung:	23. Jun 2025

कर्मण्येवाधिकारस्ते मा फलेषु कदाचन।
मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि॥

–Bhagavad Gita 2.47

You have the right to perform your duties, but not to the fruits thereof. Do not let the fruits of your actions be your motivation, nor let attachment to inaction.

Abstract

The emergence of Automated Cyber-Physical Systems (ACPS) has firmly established Autonomous Driving as a prominent application of modern technology. Despite the significant progress in this field, one of the critical barriers to the widespread adoption of Autonomous Vehicles (AVs) remains societal acceptance and trust. Public apprehension often stems from the perceived opacity of autonomous systems, leading to concerns about safety, reliability, and accountability. Addressing this challenge requires building transparency and fostering trust. One effective way to achieve this is by enabling AVs to provide clear, contextually relevant explanations of their decisions and actions.

The design of such explanations involves balancing three key dimensions: content, frequency, and timing. While prior research has made significant strides in understanding the content dimension, the other two aspects —particularly timing— remain less explored. This dissertation aims to bridge this gap by developing an algorithm that optimizes the delivery of explanations in AVs, focusing on the granularity of timing.

Existing studies on explanation timing have predominantly classified the delivery into broad categories —before, during, or after an action. For AVs, research suggests that passengers generally prefer explanations to be provided before an autonomous action is executed, as this approach aligns with their need for predictability and situational awareness. However, such broad classifications often fail to account for the nuanced interplay between cognitive load, attention, and the dynamic nature of real-world driving scenarios.

This dissertation addresses these complexities by modelling the fine-grained timing of explanations specifically tailored for AVs. The central objective is to design an algorithm capable of generating optimally timed explanations that reduce passengers' cognitive load while enhancing their trust and understanding. To achieve this, the work leverages the Salience, Effort, Expectancy, Value (SEEV) attention model, which predicts where a user's attention is likely to be focused based on the four factors of the model.

A novel aspect of this research is its exploration of multi-step explanations—explanations delivered in sequential parts rather than as a single message. The dissertation investigates how the use of multi-step explanations influences the optimal timing strat-

egy and evaluates its impact on cognitive load and user comprehension. Additionally, the study extends its scope to consider multi-user scenarios, where explanations are provided to multiple passengers or stakeholders simultaneously. The efficacy of the proposed timing strategies is evaluated through a real-world experiment conducted in a game-based setup.

In conclusion, this dissertation makes a significant contribution to the field of explainability by addressing the underexplored dimension of explanation timing in AVs. The findings highlight the importance of timing as a critical factor in designing effective human-machine interactions, demonstrating its potential to enhance transparency and trust. While the primary focus is on AVs, the insights gained from this research are broadly applicable to other domains, including healthcare, finance, and human-robot collaboration, where effective explanations play a vital role in decision-making and trust-building.

Zusammenfassung

Das Aufkommen von Automated Cyber-Physical Systems (ACPS) hat das autonome Fahren als eine herausragende Anwendung moderner Technologie fest etabliert. Trotz erheblicher Fortschritte auf diesem Gebiet bleibt eine der größten Hürden für die breite Annahme von Autonomous Vehicles (AVs) die gesellschaftliche Akzeptanz und das Vertrauen. Öffentliche Vorbehalte resultieren oft aus der wahrgenommenen Opazität autonomer Systeme, was Bedenken hinsichtlich Sicherheit, Zuverlässigkeit und Verantwortlichkeit hervorruft. Um diese Herausforderung zu bewältigen, ist es erforderlich, Transparenz aufzubauen und Vertrauen zu fördern. Ein wirksamer Ansatz hierfür ist die Befähigung von AVs, klare und kontextuell relevante Erklärungen für ihre Entscheidungen und Handlungen bereitzustellen.

Das Design solcher Erklärungen erfordert die Berücksichtigung von drei zentralen Dimensionen: Inhalt, Häufigkeit und Timing. Während frühere Forschungen bedeutende Fortschritte im Verständnis der Inhaltsdimension gemacht haben, wurden die beiden anderen Aspekte —insbesondere das Timing— weniger intensiv untersucht. Diese Dissertation zielt darauf ab, diese Lücke zu schließen, indem ein Algorithmus entwickelt wird, der die Bereitstellung von Erklärungen in AVs optimiert, mit besonderem Fokus auf die Feinheit des Timings.

Bisherige Studien zur zeitlichen Gestaltung von Erklärungen haben die Bereitstellung in breite Kategorien eingeteilt —vor, während oder nach einer Aktion. Für AVs zeigen Forschungsergebnisse, dass Passagiere in der Regel Erklärungen bevorzugen, die vor einer autonomen Aktion bereitgestellt werden, da dies ihrem Bedürfnis nach Vorhersehbarkeit und situativem Bewusstsein entspricht. Solche breiten Klassifizierungen berücksichtigen jedoch oft nicht die nuancierte Wechselwirkung zwischen kognitiver Belastung, Aufmerksamkeit und der dynamischen Natur realer Fahrszenarien.

Diese Dissertation geht auf diese Komplexitäten ein, indem sie das fein granulare Timing von Erklärungen modelliert, das speziell auf AVs zugeschnitten ist. Das Hauptziel ist es, einen Algorithmus zu entwickeln, der optimal getimte Erklärungen generiert, um die kognitive Belastung der Insassen zu minimieren und gleichzeitig ihr Vertrauen und ihr Verständnis zu fördern. Dazu wird das Salience, Effort, Expectancy, Value

(SEEV)-Aufmerksamkeitsmodell genutzt, das vorhersagt, wo sich die Aufmerksamkeit eines Nutzers wahrscheinlich konzentriert, basierend auf den vier Faktoren des Modells.

Ein neuartiger Aspekt dieser Forschung ist die Untersuchung von mehrstufigen Erklärungen – Erklärungen, die in aufeinanderfolgenden Teilen statt als einzelne Nachricht geliefert werden. Die Dissertation untersucht, wie der Einsatz von mehrstufigen Erklärungen die optimale Timing-Strategie beeinflusst und bewertet deren Auswirkungen auf die kognitive Belastung und das Nutzerverständnis. Darüber hinaus erweitert die Studie ihren Fokus auf Mehrbenutzerszenarien, in denen Erklärungen gleichzeitig an mehrere Passagiere oder Interessengruppen gerichtet werden. Die Wirksamkeit der vorgeschlagenen Timing-Strategien wird durch Experimente in einer realitätsnahen, spielbasierten Umgebung evaluiert.

Zusammenfassend leistet diese Dissertation einen bedeutenden Beitrag zum Bereich der Erklärbarkeit, indem sie sich der wenig erforschten Dimension des Timing von Erklärungen in AVs widmet. Die Ergebnisse unterstreichen die Bedeutung des Timings als einen kritischen Faktor bei der Gestaltung effektiver Mensch-Maschine-Interaktionen und zeigen dessen Potenzial zur Förderung von Transparenz und Vertrauen. Obwohl der Schwerpunkt auf AVs liegt, sind die gewonnenen Erkenntnisse auch auf andere Bereiche wie das Gesundheitswesen, die Finanzwelt und die Mensch-Roboter-Kollaboration übertragbar, in denen effektive Erklärungen eine entscheidende Rolle bei Entscheidungsprozessen und dem Aufbau von Vertrauen spielen.

Acknowledgement

This four (almost five!) year long journey wouldn't have been possible without the unwavering support of my four pillars of strength: my supervisor Martin Fränzle, my parents, my in-laws, and my constant support Constantin.

First, my deepest gratitude goes to my supervisor, Prof. Martin Fränzle, whose guidance, patience, and expertise have been invaluable throughout this process. His insights and encouragement have helped shape not only this thesis but also my academic and personal growth.

I would also like to thank Appa and Amma for believing that their child could venture into a foreign land and thrive. Your unwavering faith in me —that I can achieve anything I set my mind to— has been my greatest motivation. I wouldn't have come this far without your love and sacrifices, and for that, I am forever grateful.

To my in-laws, Astrid and Gerd, thank you for welcoming me into your family with open arms and for being like a second set of parents to me. Your kindness and warmth have ensured that I never felt alone.

And last but certainly not least, Constantin —what would I do without you? My constant cheerleader, my biggest supporter, my forever companion. From proofreading all my work to travelling to conference venues just to give me mental support, you have always been there. Now, as I take my final steps in my doctoral journey, I am happy that I can do the same for you in yours.

I am also grateful to all my former and current colleagues for the friendly work environment and countless coffee/floor chats about anything and everything. My current supervisor, Maike Schwammberger, who believed in me and my research and hired me even before I completed my PhD. A special shoutout to the M2 group, Maren and Jacob, for the constant support —especially during the early home-office days— where you unknowingly helped me navigate imposter syndrome.

It is only fair for me to give due credit to ChatGPT, which has helped me refine my ideas by improving sentence structure and clarity.

Finally, to all my friends and family who have supported me in any way during this journey — thank you! Your encouragement and belief in me have made all the difference.

Contents

1 Introduction

Contents

1.1 Motivation

In today's world, where technology is becoming more automated and autonomous, systems are evolving faster than ever before. These modern systems are not only complex but often behave in ways that older generations of devices couldn't achieve. As people interact more closely with these advanced technologies, especially those that can adapt and act on their own, understanding and predicting their behaviour becomes challenging. There's also an emerging need for systems to clearly explain their actions, as clear explanations can improve safety, build trust, and increase public acceptance of these technologies.

This dissertation focuses on finding the best way to provide explanation(s) to user(s) for actions taken by Autonomous Vehicles (AVs). Specifically, we present algorithms designed to optimize explanation timing within a game-based framework.

In fig. 1.1, we illustrate our concept of an interaction between an AV and its user in a game-like setup. In this scenario, the AV observes its environment (referred to as Env(AV)), which includes not only physical surroundings but also the human user. Based on its observations, the AV generates an explanation for the human. Meanwhile, the human receives this explanation, processes it, and observes their own environment

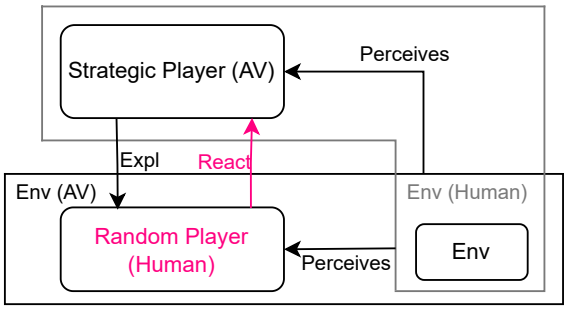


Figure 1.1: An overview of the game

(Env(Human), which consists of the surroundings as well as the AV itself) to decide on their next action.

The AV's goal in this interaction is to provide explanations at the most effective moments, aiming to prevent any increase in the human's mental workload. By delivering information at the right time, the AV can enhance user understanding without adding unnecessary cognitive effort.

1.1.1 Why use Formal Methods?

Artificial Intelligence (AI) and Fomal Methods (FM) are both used to solve complex problems in computer science, but they have different approaches, strengths, and applications.

Machine Learning (ML), neural networks, and reinforcement learning are some of the examples of AI. ML uses data-driven methods, learning from examples or experiences to make predictions, recognize patterns, or make decisions without being explicitly programmed for every situation. AI works well in unstructured, uncertain environments where data is abundant. It can adapt to new data and improve over time and is often more flexible and scalable for real-world problems like image recognition, language translation, or game playing. However, AI often lacks interpretability (i.e. commonly referred to as "black box" systems) and does not provide formal guarantees of correctness or safety. Apart from this, AI requires large datasets for training and may overfit or fail in outlier cases.

FM use mathematically rigorous techniques for the specification, verification, and validation of systems, particularly in safety-critical environments. Some of the ways of doing this are model checking, theorem proving, abstract interpretation, etc. FM provides strong guarantees of correctness, safety, and security. It is used in high-stakes domains like avionics, medical devices, and cryptography. It is also known to handle

edge cases and offers full coverage of possible system states. On the flip side, it is computationally expensive and not easily scalable for very large or complex systems. It is also typically less adaptable than AI systems, as it is often challenging to modify formal specifications. It also requires a deep understanding of the system and its environment to model accurately.

AI excels in addressing dynamic, uncertain environments where real-time adaptability is essential, while FM is crucial in safety-critical systems requiring guaranteed correctness. Rather than one being superior, AI and FM complement each other depending on the application: AI is more suited for open, data-rich systems like natural language processing or image processing, while FM ensures reliability in closed, safety-critical domains such as flight control or nuclear reactors. Hybrid approaches combining AI with formal verification are emerging, offering both adaptability and safety, potentially leading to more robust and intelligent systems. However, for our approach, FM is the preferred choice, as our test cases involve safety-critical systems, and we lack the large datasets necessary to effectively train AI models.

1.2 Research Objectives

As Autonomous Vehicles(AVs) become more prevalent, the ability of these systems to communicate effectively with human users is critical to ensuring trust, safety, and understanding. Research has shown that explanations delivered at the wrong time can overwhelm users or lead to misinterpretation, especially in dynamic environments like driving [THK24]. To address this challenge, this research focuses on the main question of *how to develop a framework for determining the optimal time to provide explanation(s)*. To aid the development of this framework following sub-research questions were generated and have been addressed in this dissertation.

1. **Developing a Framework for Optimal Atomic Explanation Timing:** The primary objective of this research is to create a framework that determines the optimal timing for delivering atomic (single-step) explanations to human users. This framework will consider the individual's attention level, enabling a more effective and adaptive explanation delivery system.

2. **Enhancing the Framework for Multi-Step Explanations:** Building on the initial framework, the second objective is to extend its capability to include multi-step explanations. This enhancement will allow the system to discern the most

appropriate moments to deliver explanations that require multiple steps, ensuring that the user's cognitive load is managed effectively throughout the process.

3. **Applying the Framework Across Multiple Users:** The final objective is to apply the refined framework in scenarios involving multiple human users. This will allow for the identification and validation of optimal timing strategies for both atomic and multi-step explanations when interacting with diverse individuals. The goal is to establish generalizable timing strategies that can be adapted for different users based on their attention profiles.

1.3 Related Work

In this section, we provide an overview of related work relevant to the overall content of the thesis. For the specific content discussed in the following chapters, we include dedicated reviews of related work within the respective chapters and sections. This related work section mainly deals with the following two aspects:

1. *Explanation models:* A discussion of existing explanation models and their various dimensions.

2. *Cognition in explanation:* An exploration of the role cognition plays in shaping and understanding explanations.

1.3.1 Explanation Models

Lewis defines explanation as "to explain an event is to provide some information about its causal history" [Lew86, p.217]. According to Walton, a successful explanation occurs when understanding is effectively transferred between the explainer and the addressee [Wal04]. However, Markus et al. [MKR21] and Ferrario & Loi [FL22] claim that explanations not only improve comprehension but are also commonly connected to higher levels of trust in automated systems.

Explanation models provide frameworks for autonomous systems to communicate their decision-making processes, helping users understand, trust, and interact safely with these technologies. As autonomous systems ——ranging from AI-driven software to Autonomous Vehicles (AVs)—— integrate into daily life, effective explanations become vital for managing user expectations and promoting safe, trusted use. This section reviews key categories of explanation models, drawing from recent research to illustrate their core characteristics and applications.

Content-Based Models

Content-based models focus on the substance of explanations, addressing fundamental questions like "*What is happening?*" and "*Why is it happening?*" [Mil19]. By determining the most relevant details to communicate, content-based models aim to balance informativeness with cognitive simplicity, ensuring that explanations are neither too complex nor too vague. For instance, Hoffman et al. [Hof+18] suggest that the right level of detail is critical for maintaining user engagement and trust.

These models are particularly useful in scenarios where users need a basic understanding of a system's actions without overwhelming technical detail, such as in AV navigation systems where passengers may need quick reassurance about vehicle actions. Context-based explanations can be expressed in different ways. Following are two of the ways.

- **Causal and Counterfactual Explanation Models:** Causal models focus on explaining cause-and-effect relationships, providing users with insights into the factors influencing a system's actions [PM18]. This approach is especially valuable in complex decision-making environments where transparency is crucial, such as healthcare or finance [RSG16]. Counterfactual models, in contrast, explain what would happen if different choices or actions were taken, effectively answering "what if" questions [Gui22]. These models help users understand the system's flexibility and adaptability by highlighting how it would respond under alternative scenarios [WMR17]. Both causal and counterfactual explanations have been shown to increase user confidence by elucidating the underlying logic of a system's actions [Sch+20; WKB22; WBK23].

- **Transparency and Justification Models:** Transparency and justification models aim to make a system's processes more visible and justifiable by explaining decision-making rules, ethical considerations, or compliance with regulations [WS21; LFV21]. These models are particularly relevant for sensitive applications where accountability is paramount. For instance, AVs may use justification models to reassure passengers by explaining safety protocols during high-risk manoeuvres [Hof+18]. Justification models also support compliance with emerging ethical and legal standards, providing stakeholders with the confidence that the system adheres to accepted practices and guidelines [Bin+18].

Timing and Frequency Models

Timing and frequency models determine when and how often explanations should be delivered to avoid cognitive overload [DK17; ZYR21]. Research indicates that timing plays a crucial role in managing users' mental workload, with pre-action explanations often being more effective than post-action explanations in high-stakes applications like AVs [Koo+16; Has+18; Du+19]. Timing is also an important factor in influencing the trust of humans in autonomous systems [Mar+17; Ros+20]. Frequency control is also essential; delivering explanations too frequently can lead to redundancy and mental fatigue, especially as users become more familiar with the system [Kul+13a; Sch+19]. Shen et al. also observed that users require explanations primarily in specific scenarios, particularly when they encounter near-death situations [She+20]. Optimal timing and frequency help reduce cognitive strain, ensuring explanations support rather than hinder user understanding [RTC18]. Körber et al. [KPB18] also found that providing an explanation at a later point (14 seconds) after asking drivers to take-over improved their understanding of the situation. Chen et al. investigated how the timing of explanations (pre, post, both, or none) influences user trust, understanding, and satisfaction with AI systems [CLS24]. Their findings reveal that pre-explanations are more effective for biased AI, post-explanations work better for unbiased AI, and providing both enhances trust calibration. Kim et al. suggest that well-designed explanations delivered at the right moments can help passengers build appropriate trust in automated vehicles by providing transparency about the vehicle's current state and actions [Kim+24].

User-based Models

User-based models modify explanations based on the user's expertise, preferences, and context [BCM07]. Using data on users' past interactions, these models tailor explanations dynamically, providing more comprehensive explanations for new users and brief updates for experienced users. For example, Hayes and Shah [HS17] highlight that personalized explanations can improve user satisfaction and decrease the cognitive load by adjusting to individual needs. By personalizing explanations, user-based models can significantly enhance user trust and engagement over repeated interactions with the system [Gun+21; Pap+23]. Prior work has shown that explanations can enhance users' mental models of a system [Chi09]. Wiegand et al. [Wie+19] evaluated the information drivers need explained during unexpected AV behaviour, resulting in a target mental model that combines key elements of expert and user mental models. Schwammberger

and Klös, in their work [SK22], propose a process for extracting explanation models from system specification models and refining them for specific users and situations.

Self-Explaining Models

In the last few years, efforts have been made to develop conceptual frameworks that offer a structured, process-oriented approach to generating explanations. This need to include self-explainability in AVs, and autonomous systems in general, arises due to the IEEE Standard for Transparency of Autonomous Systems, which emphasizes the need for making autonomous systems understandable to the stakeholders [22]. In 2019, Blumreiter et al. [Blu+19] introduced the MAB-EX framework (Manage, Analyze, Build, Explain), drawing inspiration from the self-adaptive MAPE loop (Monitor, Analyze, Plan, Execute) [Sin06]. MAB-EX systematically guides the explanation generation process by focusing on managing information, analyzing system behaviour, building insights, and delivering explanations tailored to user needs, thereby enhancing understanding and supporting decision-making. Similarly, Ziesche et al. [ZKG21] propose a method for detecting and classifying anomalous behaviour in autonomous systems to enable self-explainability. This approach allows systems to autonomously explain behaviours that deviate from anticipated outcomes, improving transparency and trust in autonomous operations.

Fey et al. [FFD22] present a framework for self-explaining systems that engage in interactive communication to clarify their operations to an addressee. The framework uses a generic explanation pattern that supports various explanation types ——causal, counterfactual, abstract, example-based, and strategy-exposing—— while establishing shared terminology between the system (E) and the addressee (A). Both parties rely on local models (M_E and M_A) and beliefs (B_E and B_A) about the world, incorporating state information and dynamics to infer future outcomes. By forming a belief ($B_E(A)$) about the addressee's understanding, E tailors explanations to bridge the gap between the addressee's interpretation and the actual situation, ensuring situational relevance and alignment with the addressee's comprehension.

1.3.2 Cognition in Explanation

In the previous subsection, we have discussed the various explanation models that exist and their potential to improve trust in AVs while enhancing the understanding of the decision-making steps taken by these systems. The recipients of these explanations can either be other AVs or Human Agents (HAs). This thesis focuses on the latter,

emphasizing the unique challenges of designing explainability processes for human recipients [Lan+21]. Effective explainability depends significantly on cognitive models, which describe how humans perceive and interpret information. Despite their complexity, these models are crucial for integrating safety mechanisms grounded in formal tools and methodologies [Car+24; Xu+21]. The design of these cognitive models varies greatly, as human behaviour and understanding are influenced by numerous psychological and situational factors [Vis06]. These models play a pivotal role in bridging the gap between the technical logic of AVs and the mental models of HAs.

Cognitive models have played a critical role in understanding human cognition, including reasoning, problem-solving, and decision-making processes. Early foundational work by [SN71] introduced cognitive architectures like the General Problem Solver (GPS), which formalized the strategies employed by humans in problem-solving scenarios. This laid the groundwork for further developments in cognitive science, particularly with models such as ACT-R ([And96]), which integrates various cognitive processes including memory and learning.

Theories of mental models have greatly enhanced our understanding of cognitive processes. Johnson-Laird [Joh86] proposed that individuals create mental representations of the world, which serve as a foundation for reasoning and inference. These mental models help explain how people generate conclusions and make judgments, even when faced with limited information. However, human cognitive resources—such as memory, attention, and information-processing capacity—are inherently limited, which can impede the ability to comprehend complex explanations [Swe88]. Additionally, individuals tend to focus selectively on information they deem most relevant. As a result, for an explanation to be effective, it must align closely with the recipient's immediate goals and priorities [Kah73].

Recent advancements in cognitive modelling have emphasized the importance of incorporating social and contextual factors into cognitive theories. For instance, research by [Mil19] highlights how explanations are shaped by social interactions and the cognitive biases that affect human reasoning. This perspective suggests that cognitive models must consider not only individual cognitive processes but also the broader social dynamics that influence understanding.

Additionally, work by [Pea09] on causal inference has underscored the significance of causal reasoning in cognition. Causal Bayesian Networks have been employed to model how individuals draw conclusions based on observed relationships, providing insights into human reasoning that can inform cognitive models.

In cognitive science, explanation is seen as a fundamental mental activity that allows individuals to make sense of the world by organizing knowledge and identifying causal relationships. Early theories, such as those by Piaget [Pia05], suggested that humans have an intrinsic drive to seek coherence and balance in their understanding of their environment, leading to the generation of explanations as a way to resolve cognitive dissonance. Tversky and Kahneman's work [TK74] on heuristics and biases further highlighted the idea that people often use simplified cognitive shortcuts to generate explanations, even when these are imperfect or incomplete.

More recently, Lombrozo's work on explanation emphasized the role of explanatory virtues such as simplicity, generality, and coherence in shaping how people evaluate explanations [Lom16]. Her research shows that people prefer explanations that are simple yet powerful, striking a balance between explanatory depth and cognitive economy. Cognitive models derived from her work have helped clarify how individuals prioritize certain types of explanations, revealing an interplay between explanatory preferences and cognitive load.

In decision-making, explanations are vital for justifying choices and actions. Johnson-Laird's mental models theory posits that people create mental simulations to reason about potential outcomes, and explanations help refine these models by focusing attention on causal mechanisms [Joh86]. Cognitive models of explanation in decision-making thus emphasize causal reasoning and counterfactual thinking, where individuals consider "what-if" scenarios to generate explanations that help them navigate uncertainty and make informed choices.

The growing interest in explainability in AI has led to the development of computational models inspired by human cognitive processes. Cognitive models of explanation have informed the design of Explainable AI (XAI) systems, which aim to make machine reasoning more transparent and understandable to human users. Miller provides a comprehensive review of how cognitive science theories, particularly those related to causal reasoning, have influenced the development of XAI systems [Mil19]. He argues that for AI explanations to be effective, they must align with human expectations of causality, coherence, and simplicity.

While cognitive models have provided valuable insights into explanation, several challenges remain. One key issue is the variability in how different individuals generate and evaluate explanations. Cognitive models often assume a normative framework for explanation evaluation, but individual differences in cognitive styles, background knowledge, and cultural factors can lead to divergent preferences for certain types of explanations.

The Role of Emotion in Explanatory Processing Cognitive models traditionally focus on the rational aspects of explanation, such as causal reasoning and coherence. However, emotions also play a significant role in how people seek, generate, and evaluate explanations. For example, explanations that evoke trust or reduce anxiety may be more readily accepted, even if they are not the most logically sound.

In this work, we incorporate cognitive models to account for the profound impact of explanations on human cognition. Understanding how cognition influences the interpretation of explanations allows us to design communication strategies that align with human mental processes. This alignment is particularly critical in determining the optimal timing for providing explanations, ensuring they are delivered in a manner that supports comprehension and minimizes cognitive overload.

1.4 Contributions and Publications

The existing works related to explanation content and timing are summarised in this section. As we can see in the fig. 1.2, on the explanation content axis, we talk about who the recipients of the explanations are, papers that ask "what and why" questions, and how these explanations are provided. On the explanation timing axis, we talk about no/generic timing, before an action is performed, during an action, and after an action is performed.

The existing papers discuss timing in a largely abstract manner, without delving into the finer nuances of when explanations should be provided. There is little to no research on predicting the optimal timing for delivering explanations ——either before or during an action—— particularly in contexts involving consistent content, which could serve as a baseline for analysing different types of explanations. This dissertation addresses this gap by focusing specifically on this aspect, supported by the publication of the following papers by the author.

1. Akhila Bairy. "Modeling Explanations in Autonomous Vehicles." In: *Integrated Formal Methods - 17th International Conference, IFM 2022, Lugano, Switzerland, June 7-10, 2022, Proceedings*. Ed. by Maurice H. ter Beek and Rosemary Monahan. Vol. 13274. Lecture Notes in Computer Science. Springer, 2022, pp. 347–351. DOI: 10.1007/978-3-031-07727-2_20. URL: https://doi.org/10.1007/978-3-031-07727-2%5C_20

2. Akhila Bairy, Willem Hagemann, Astrid Rakow, and Maike Schwammberger. "Towards Formal Concepts for Explanation Timing and Justifications." In: *30th IEEE*

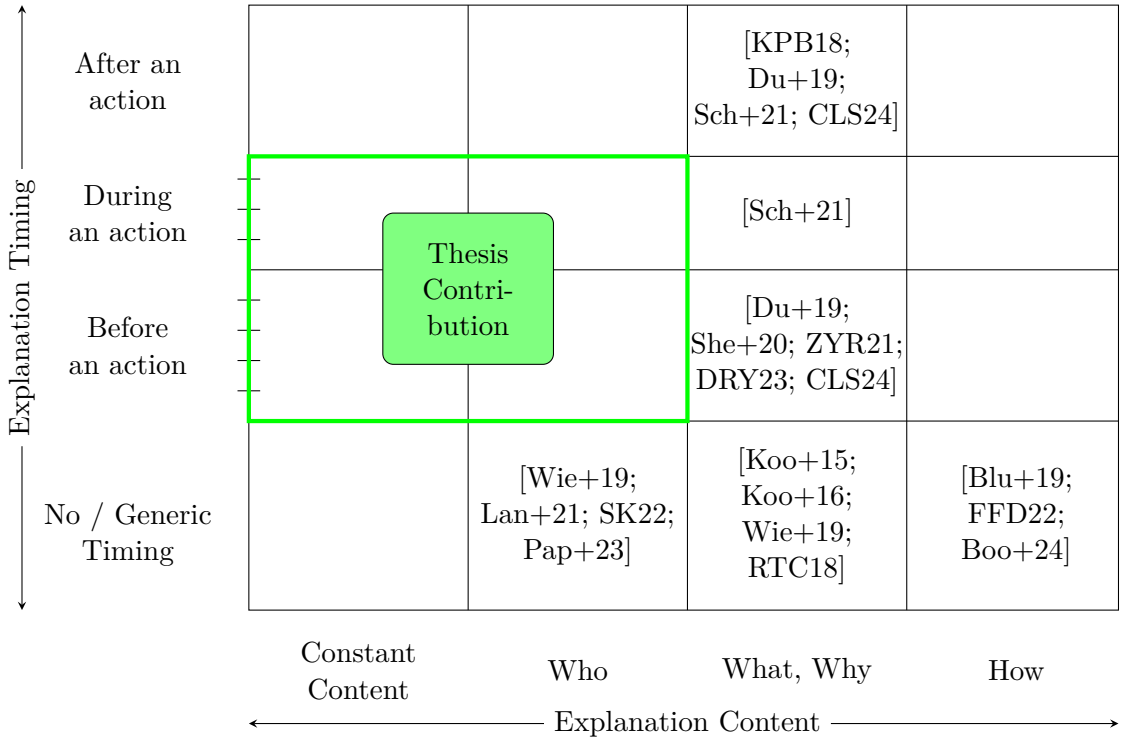


Figure 1.2: Existing works related to explainability

International Requirements Engineering Conference Workshops, RE 2022 - Workshops, Melbourne, Australia, August 15-19, 2022. IEEE, 2022, pp. 98–102. DOI: 10.1109/REW56159.2022.00025. URL: https://doi.org/10.1109/REW56159.2022.00025

3. 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

4. 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

5. 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

6. Akhila Bairy and Martin Fränzle. "What if Autonomous Systems had a Game Master? Targeted Explaining with the help of a Supervisory Control System." In: ExEn '24. Lisbon, Portugal: Association for Computing Machinery, 2024, pp. 15–19. ISBN: 9798400705960. DOI: 10.1145/3648505.3648508. URL: https://doi.org/10.1145/3648505.3648508

7. Mehrnoush Hajnorouzi, Astrid Rakow, Akhila Bairy, Jan-Patrick Osterloh, and Martin Fränzle. "What Level of Power Should We Give an Automation?" In: *Dependable Computing – EDCC 2024 Workshops*. Ed. by Behrooz Sangchoolie, Rasmus Adler, Richard Hawkins, Philipp Schleiss, Alessia Arteconi, and Adriano Mancini. Cham: Springer Nature Switzerland, 2024, pp. 14–21. ISBN: 978-3-031-56776-6. DOI: 10.1007/978-3-031-56776-6_2

8. Akhila Bairy and Martin Fränzle. "Enhancing Multi-user Experience: Optimizing Explanation Timing Through Game Theory." In: *Intelligent Technology for Future Transportation*. Ed. by Abolhassan Razminia and Dinh Hoa Nguyen. Cham: Springer Nature Switzerland, 2025, pp. 106–117. ISBN: 978-3-031-84148-4

9. Maike Schwammberger, Astrid Rakow, Lina Putze, and Akhila Bairy. "Explain it for Safety: Explanations for Risk Mitigation." In: *Design and Verification of Cyber-Physical Systems: From Theory to Applications*. Ed. by Andreas Rauh, Bernd Finkbeiner, and Paul Kröger. to be published. 2025

10. Akhila Bairy, Martin Fränzle, and Maike Schwammberger. *Optimising Timing of Multi-Step Explanations for Multiple Users using Reactive Game*. Accepted to the 7th International Workshop on EXplainable, Trustworthy, and Responsible AI and Multi-Agent Systems (EXTRAAMAS 2025). 2025

11. Akhila Bairy, Mehrnoush Hajnorouzi, Astrid Rakow, Martin Fränzle, and Maike Schwammberger. "Timing Matters - A Study on the Role of Timing in Explanation Delivery." In: *Human Systems Engineering and Design (IHSED2025): Future Trends and Applications* 198 (2025). DOI: http://doi.org/10.54941/ahfe1006782

1.5 Structure of the Thesis

This thesis is structured into several chapters, each building on the foundational concepts and progressively addressing more complex issues related to explanation optimization for single and multiple users in real-world scenarios. Here is an overview of each chapter's focus and contribution to the thesis as a whole:

In chapter 2, we introduce the fundamental concepts that provide the theoretical basis for this research. This includes a detailed overview of Game Theory, which is essential for understanding strategic decision-making processes, as well as cognitive models that explain how users interpret and process information. Together, these frameworks serve as the foundation for the subsequent chapters.

Chapter 3 and chapter 4 tackle the optimization problem for explanations directed at a single user. Chapter 3 focuses on optimizing single-step/atomic explanations, aiming to maximize clarity and understanding in situations where brief, immediate information is needed. Chapter 4 extends this by exploring multi-step explanations, which are necessary when information must be conveyed over a sequence of interactions to build a more comprehensive understanding.

However, real-world scenarios often involve multiple individuals (passengers) in a car. This complexity introduces the need for adaptive models that cater to multiple users simultaneously. Chapter 5 and chapter 6 address this challenge by adapting the single-user model to multi-user settings. In chapter 5, we develop methods for optimizing the timing of atomic explanation for multiple users. Chapter 6 extends these strategies to multi-step explanations for multiple users.

In chapter 7, we present a user study aimed at investigating the timing of explanations and their impact on users' understanding. This chapter explores how the timing of information delivery affects comprehension and user satisfaction, providing empirical evidence to support our theoretical models.

Finally, the thesis concludes with chapter 8, where we summarize the findings and contributions of this research. This chapter synthesizes the insights gained from each preceding chapter, highlighting the advancements made in optimizing explanations for both single and multi-user scenarios. Additionally, it outlines the limitations of the study and explores potential directions for future research.

2 Fundamentals

Contents

In this chapter, we introduce the fundamental concepts used in our work. Our work includes a reactive game, and hence in section 2.1, we introduce game theory and discuss some of the different forms of games. Followed by an overview of Backward Induction in section 2.2, which is used in the game. We started the implementation of the reactive game with PRISM Model Checker and then shifted to MATLAB. Therefore, in section 2.3, we talk about the PRISM Model Checker and its drawbacks, and in section 2.4, we introduce MATLAB and explain why we use it. Another major aspect of this work is the use of an attention model. In section 2.5, we give a brief overview of different cognitive models that exist, followed by a focus on the SEEV attention model in the section 2.6.

2.1 Game Theory

Game Theory, pioneered by mathematicians John von Neumann and Oskar Morgenstern in their seminal work Theory of Games and Economic Behaviour (1944) [NM07], is a mathematical framework for analyzing strategic interactions among rational agents. It provides tools to model and study situations where the outcome for each participant depends not only on their own decisions but also on the decisions of others. Since its inception, game theory has evolved into a multidisciplinary field, with applications in economics, political science, biology, computer science, and psychology.

At its core, Game Theory addresses scenarios involving conflict, cooperation, and competition. The participants, referred to as *players*, make decisions based on their preferences, strategies, and expectations of others' behaviour. Each player aims to maximize their *payoff*, which reflects their satisfaction or utility from a particular outcome [OR94].

Game Theory distinguishes between different types of games based on the structure of interaction:

- *Players and Strategies:* Players are decision-makers, and strategies are their particular choice of actions. A player's choice of strategy can depend on their knowledge of the game and their expectations about others' choices [FT91].
- *Payoffs:* Each combination of strategies results in a specific outcome, which corresponds to a payoff for each player. These payoffs can be represented numerically, capturing the players' preferences over outcomes.
- *Equilibria:* A central focus of Game Theory is identifying equilibria, where players' strategies are mutually consistent. The most notable is the Nash equilibrium, where no player can unilaterally improve their payoff by changing their strategy [Nas50].

Apart from the above-mentioned form, Game Theory can be categorised in several other ways as well. Some of the common ways to categorise is:

- *Static and dynamic games* - In static games, players make decisions simultaneously without knowing other players' choices. In dynamic games, decisions are made sequentially, allowing players to adjust strategies based on previous moves.
- *Cooperative and non-cooperative games* - Cooperative games allow players to form binding agreements and work together for mutual benefit. In non-cooperative games, players act independently, leading to competitive decision-making.

- *Zero-sum and non-zero-sum games* - In zero-sum games, one player's gain equals another's loss. Non-zero-sum games allow mutual benefit or loss, making cooperation possible.
- *Symmetric and asymmetric information games* - Symmetric games provide all players with the same information. In asymmetric games, some players have more knowledge, leading to strategic uncertainty.
- *Single-shot and iterated games* - Single-shot games involve a one-time interaction with no future consequences. Iterated games involve repeated interactions, encouraging long-term strategies and cooperation.

In Game Theory, there are two fundamental types of games —*single-shot games* and *iterated games*— that represent distinct settings in which these interactions occur. Understanding the distinction between these two types is crucial, as it significantly influences players' decision-making processes, strategies, and outcomes.

2.1.1 Single-Shot Games

Single-shot games represent the simplest type of strategic interaction. In these games, players interact only once, without any opportunity for future encounters or retaliation. As a result, players typically focus on maximizing their immediate payoff, without regard to potential future consequences.

A single-shot game is a scenario where:

- Players make decisions simultaneously or sequentially.
- The game ends after one round, with no repetition.
- Players have no knowledge of future interactions, leading them to base decisions solely on the current situation.

Formally, a single-shot game can be represented by:

- A set of players $N = \{1, 2, \ldots, n\}$,
- A strategy space S_i for each player $i \in N$,
- A payoff function $u_i(s_1, s_2, \ldots, s_n)$ for each player, where s_i represents player i's strategy.

One of the well-known examples of a single-shot game is the *Prisoner's Dilemma*. This game consists of two players (prisoners), who must independently decide whether to *cooperate* (stay silent) or *defect* (betray the other). The dominant strategy for each player is to defect, even though mutual cooperation would lead to a better collective outcome. Some of the other examples include the *Battle of the Sexes* and *Matching Pennies*.

In single-shot games, the concept of *Nash equilibrium* is often used to analyze the outcome. A Nash equilibrium occurs when no player has an incentive to unilaterally change their strategy, given the strategies chosen by the others. Since players do not expect future interaction, the strategies tend to be static and focused on immediate gain.

2.1.2 Iterated Games

In contrast to single-shot games, *iterated games* involve multiple rounds of interaction between the same players. The repeated nature of the game allows for more complex strategies, such as retaliation, cooperation, and forgiveness. The potential for future interactions influences the players' behaviour, as they may consider both short-term payoffs and long-term benefits.

An iterated game involves:

- A set of players $N = \{1, 2, \ldots, n\}$,
- A strategy space S_i^t for each player $i \in N$ in each round t,
- A sequence of payoff functions $u_i(s_1^t, s_2^t, \ldots, s_n^t)$ for each round t,
- Players receive feedback after each round and can adjust their strategies accordingly.

An iterated game can be thought of as a repeated version of a single-shot game, where the same strategic choices are presented multiple times, allowing players to adapt based on their experience and expectations of future play.

An iterated game can be classified into two types:

- *Finite games:* In this, all the players involved know that the game will be played for a limited number of rounds, after which the interaction ends.
- *Infinite games:* The players involved have no knowledge of when the game will end, meaning the interactions can continue indefinitely.

A prominent example of iterated games is the *Iterated Prisoner's Dilemma*. In this version of the classic game, all the players repeatedly choose whether to cooperate or defect over many rounds. Strategies such as *Tit-for-Tat* —where a player mimics their opponent's previous move— can emerge, fostering cooperation in the long term. Some of the other examples include the *repeated Battle of the Sexes* and *repeated coordination games*.

The possibility of repeated interactions introduces new strategic elements such as reciprocity, punishment, and reputation. Players may cooperate to build trust, knowing that defection could lead to future punishment or loss of trust. Conversely, players may defect if they believe the game will end soon, prioritizing short-term gains.

In iterated games, long-term strategies often emerge, and concepts like *subgame perfect equilibrium* become relevant. Unlike Nash equilibrium in single-shot games, a subgame perfect equilibrium ensures that players' strategies form a Nash equilibrium in every subgame (i.e., every round), allowing for more consistent and sustainable strategies across multiple interactions.

2.1.3 Single-Shot Games vs Iterated Games

In the previous two subsections, we discussed about the two forms of games: *Single-shot game* and *Iterated game*. In this subsection, let us compare the similarities and differences between single-shot game and iterated game.

Aspect	Single-Shot Game	Iterated Game
Interaction Frequency	One-time interaction	Repeated interactions over multiple rounds
Strategic Focus	Immediate payoff	Long-term payoff and future interactions
Cooperation	Less likely due to no future encounters	More likely due to potential for reciprocity
Punishment/Retaliation	Not applicable	Retaliation and punishment possible over rounds
Equilibrium Concept	Nash equilibrium	Subgame perfect equilibrium

Table 2.1: Comparison between single-shot game and iterated game

The table 2.1 gives us a brief overview of the major differences between single-shot and iterated games. The primary difference between single-shot and iterated games lies in the time horizon of the interaction. Single-shot games are typically analyzed with static strategies, while iterated games allow for dynamic strategies that adapt based on

the players' past behaviour. In single-shot games, players interact only once, focusing on immediate payoffs without any expectation of future consequences. This often leads to decisions driven by self-interest, such as defection in the Prisoner's Dilemma, where mutual cooperation could have been more beneficial. Strategic behaviour is typically simpler, as players aim to optimize their payoffs for a single round.

In contrast, iterated games involve repeated interactions, where players recognize that their actions today influence future outcomes. This creates opportunities for cooperation, as the possibility of retaliation or reward in subsequent rounds encourages strategies like *Tit-for-Tat* (mirroring the opponent's previous move) or *Grim Trigger* (maintaining cooperation unless the opponent defects). These strategies rely on the possibility of future retaliation or reward, fostering more complex and cooperative equilibria that are less common in single-shot games.

Single-shot and iterated games offer distinct perspectives on strategic decision-making. While single-shot games prioritize immediate outcomes, encouraging simpler, self-serving strategies due to the absence of future consequences, iterated games emphasize the ongoing nature of interactions, where the potential for future rewards or retaliation fosters the development of more sophisticated strategies and sustained cooperation over time.

2.1.4 Reactive Games

Reactive games form an important subset of game theory, where players continuously adjust their strategies based on the observed actions of others [MN19]. Unlike static or sequential games, where strategies are pre-determined or follow a structured order, reactive games emphasize dynamic decision-making and adaptation.

A defining feature of reactive games is the continuous strategy adjustment, where players modify their actions in response to the behaviour of their opponents. This ongoing interaction creates a feedback loop, influencing the strategic decisions of all participants. For instance, in iterated games, players often employ reactive strategies that depend on the opponent's previous moves, leading to complex patterns of cooperation and competition. These games are commonly modelled using Markov Decision Processes (MDPs) and reinforcement learning frameworks, where strategies evolve dynamically in response to new information [Lit94].

Theoretical exploration of reactive games has led to the development of various models and strategies. One notable approach involves reactive learning strategies, which gradually adjust a player's propensity to take certain actions based on past interactions with opponents. These strategies have been shown to effectively restrict the set

of feasible payoffs in iterated games, highlighting their potential in influencing game outcomes [MN19].

Another significant concept is the reactive bargaining set, which examines the stability and dynamics of agreements in cooperative games. This framework has been instrumental in understanding how players can reach and maintain mutually beneficial agreements in environments where strategies are continuously evolving [GM97].

In practical applications, reactive game theory has been utilized in the synthesis of autonomous systems. For example, in scenarios involving information asymmetry, opportunistic synthesis methods have been developed to enable autonomous agents to achieve better outcomes by leveraging incomplete information. This approach allows for the construction of control systems that can adapt to dynamic and unpredictable environments [KF19].

2.1.5 Formal Definition of a Game

In game theory, a **game** is defined as a mathematical structure consisting of several key components that describe the strategic interaction among rational decision-makers, called players. A generalised formalisation of a game $\mathcal{G}$ can be represented as a tuple:

$$\mathcal{G} = (\mathcal{P}, \mathcal{S}, \mathcal{A}, T, R),$$

where:

1. **Players ($\mathcal{P}$)**: The finite set $\mathcal{P} = \{1, 2, \ldots, n\}$ represents the players in the game. Each player $i \in \mathcal{P}$:
   - Is a rational agent that seeks to optimize their individual objective (e.g., maximizing rewards or minimizing costs).
   - Can interact with the environment (via actions) and observe changes in the game state.
   - May adopt strategies that are influenced by other players' actions or states, depending on the game type (e.g., cooperative, competitive, or mixed).
2. **States ($\mathcal{S}$)**: The set $\mathcal{S}$ contains all possible configurations or situations in the game. Each state $s \in \mathcal{S}$:
   - Encodes the current environment and possibly players' private information (in incomplete information games).

- May include static elements (e.g., board positions, resources) and dynamic elements (e.g., time, player statuses).
- Can be deterministic or stochastic, depending on the game dynamics and transitions.

3. **Actions ($\mathcal{A}$)**: For each player i, $\mathcal{A}_i$ represents the set of actions they can choose. The joint action space is:

$$\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n,$$

where:

- $a_i \in \mathcal{A}_i$ is an action chosen by player i.
- $a = (a_1, a_2, \ldots, a_n) \in \mathcal{A}$ is the joint action profile, representing the combined actions of all players.
- Actions may be discrete (e.g., move left or right), continuous (e.g., adjust speed), or mixed.

4. **Transition Function (T)**: The transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ describes how the game state evolves:
   - Given the current state $s \in \mathcal{S}$ and joint action $a \in \mathcal{A}$, the next state $s' \in \mathcal{S}$ is determined as $T(s, a) = s'$.
   - Transitions may be:
     - *Deterministic*: The next state is uniquely determined by s and a.
     - *Stochastic*: The next state is drawn from a probability distribution $P(s'|s, a)$.

5. **Reward Function (R)**: The reward function $R = (R_1, R_2, \ldots, R_n)$ assigns immediate feedback to players based on the state and actions:
   - Each $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ maps a state $s \in \mathcal{S}$ and joint action $a \in \mathcal{A}$ to a real-valued reward $R_i(s, a)$ for player i.
   - Rewards can represent:
     - Points scored in a game.
     - Costs incurred (e.g., penalties or resource usage).
     - Progress toward a goal.
   - The reward function guides players' decision-making and may reflect competitive or cooperative objectives.

2.1.6 Timed Games

Timed games are a subclass of games played on timed automata, which extend finite-state machines by incorporating clocks that track the passage of time. Timed games are often employed in various fields such as cognitive psychology, Human-Computer Interaction (HCI), and behavioural research [ABG10; BIF15; Rie+23; RZG14]. They are designed to assess how participants perform under time constraints. These games require the players to complete tasks within a specified time limit or as quickly as possible.

A common objective in timed games is to find a strategy that ensures a player's goal is met within specific time constraints, often expressed as reachability, safety, or optimization problems [AD94].

We can formally define a timed game $\mathcal{G}$ as a tuple:

$$\mathcal{G} = (\mathcal{S}, \mathcal{A}, T, \delta, C, R, \Phi),$$

where:

- $\mathcal{S}$: The set of states representing all possible configurations of the game.
- $\mathcal{A}$: The set of actions available to players.
- T: The time domain, typically continuous ($\mathbb{R}_{\geq 0}$) or discrete ($\mathbb{N}$).
- C : Clocks $\to \mathbb{R}_{\geq 0}$: A set of clock variables tracking elapsed time for transitions and constraints.
- $\delta : S \times A \times (C \to \mathbb{R}_{\geq 0}) \times T \to S$: The transition function specifies how the system evolves based on the current state, the chosen action, the current clock valuations, and the elapsed time.
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: A reward function that assigns a numerical value (e.g., cost or utility) to state-action pairs, representing objectives such as minimizing cost or maximizing payoff over time.
- Φ: Temporal objectives or winning conditions, often expressed using temporal logic, such as reachability, safety, or optimization within given time constraints.

2.1.7 Markov Decision Game

Markov Decision Process (MDPs), originated in the 1950s and first studied extensively in the 1960s, provide a robust framework for solving dynamic decision-making problems in stochastic environments [Bel57; How60]. These procedures are widely known as

stochastic dynamic programming, emphasizing their ability to address decisions across multiple periods under uncertainty.

A Markov Decision Game, also known as 1.5-player game, developed by Lloyd Shapley in the 1950s [Sha53], is a theoretical framework that is based on Markov Decision Process (MDP) involving interaction between two distinct types of agents: a fully controllable player, aka a strategic player, and a partially observable adversary, aka a random player. This concept is used to model decision-making problems where one of the agents (the strategic player) has complete control over their actions, while the second "half" player represents uncertainties or stochastic elements of the environment, which influence the outcome but are not directly controlled by any strategic player. The primary goal is for the strategic player to make optimal decisions under uncertainty, considering both the controllable actions and the probabilistic nature of the adversarial environment [SV15].

In a 1.5-Player Game, a fully observable agent (strategic player) interacts with a probabilistic system, often modelled as a random player. The random player represents environmental randomness or uncontrollable adversarial actions. Unlike two-player games, the second player in a 1.5-player game is not an active decision-maker but a source of uncertainty [Put94]. Apart from that, in a Markov Decision Game, the system state transitions depend only on the current state and the action taken by the strategic player, following the Markov property. The transition probabilities can be influenced by the probabilistic nature of the half-player [Put94].

The goal of the strategic player is to optimize a performance criterion, such as expected reward, over time by accounting for both the deterministic actions they can control and the stochastic effects introduced by the half-player.

Formal Definition of a Markov Decision Game

Formally, a Markov Decision Game is defined by the tuple (S, A, P, R, γ), where:

- S: represents a finite set of states representing the different configurations of the game environment.
- A: represents a finite set of actions available to the main/strategic player. Each action impacts the state transitions within the environment.
- $P(s'|s, a)$: is the state transition probability function, which specifies the probability of moving from state s to state s' given that action a is chosen.

- $R(s, a)$: represents a reward function that assigns a real-valued reward based on the state-action pair (s, a), representing the utility the player receives for taking action a in the state s.
- $\gamma \in [0, 1]$: represents a discount factor that weights future rewards relative to immediate rewards. The factor controls how much importance is given to long-term versus short-term gains.

The objective of the strategic player in a Markov Decision Game is to find a policy $\pi : S \to A$ that maximizes the expected cumulative reward (or value function $V^{\pi}(s)$) over time, starting from any given initial state s. The value function under policy π is defined as:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_t = \pi(s_t)\right] \tag{2.1}$$

where s_t represents the state at time t, and a_t is the action selected by the policy π in state s_t.

2.2 Backward Induction

Backward Induction (BI), also called Backward Bellman Induction, is a mathematical technique, often used in Game Theory, for solving finite extensive-form games. BI involves working backwards from the end of a sequence of decisions or events to determine the optimal decision at each step [Hei12]. It is used to derive optimal strategies in finite games where decisions are made in stages, and where each player prepares for any reasonable action of others, and deduces the sequence of actions that maximizes payoffs at each stage. The technique works by iterating backwards from the final stage of the game to determine the best possible action at every earlier point, assuming that all participants are rational and seek to maximize their payoffs. The first known usage of BI was by Arthur Cayley in 1875 where he used this method to try and solve the secretary problem [Rus16].

The concept of backward induction builds on the principle of subgame perfect equilibrium (SPE), which is a refinement of Nash equilibrium for extensive-form games [FT91]. In a game with multiple decision points, BI ensures that each subgame—an independent part of the overall game—reaches an optimal outcome. This is achieved by first solving the final subgame and then proceeding to earlier subgames in reverse order.

At each decision node, BI assumes that the player making the choice will select the action that maximizes their payoff, given the subsequent moves of other players. By working backwards from the final outcome, backward induction eliminates non-credible threats and strategies, focusing only on those actions that players would logically pursue under rational behaviour. The process results in a strategy profile that constitutes a subgame perfect equilibrium, ensuring optimal decision-making at each stage of the game.

Backward Induction is especially useful in the analysis of sequential games with perfect information, where players make decisions one after another, fully aware of previous actions. It has been successfully applied in various domains, including economics, political science, and negotiation theory.

While Backward Induction is a powerful method, its reliance on the assumption of rationality may limit its applicability in real-world scenarios where agents may exhibit bounded rationality or irrational behaviour. Additionally, in games with incomplete or imperfect information, BI may not be as effective, since players may not fully anticipate the moves of others due to uncertainty or information asymmetry.

Although Backward Induction remains a foundational tool in the analysis of sequential decision-making processes, offering clear insights into optimal strategies in settings where future actions can be predicted with a high degree of certainty, its practical utility may be constrained by the complexity and unpredictability of human behaviour in non-ideal conditions.

2.3 PRISM Model Checker

The PRISM Model Checker is a formal verification tool used to model and analyze systems that exhibit probabilistic behaviour [KNP11]. It is widely used in the verification of complex systems like network protocols, biological systems, and embedded systems where uncertainty and randomization play a significant role.

PRISM supports various mathematical models tailored to different types of systems. For instance, Discrete-Time Markov Chains (DTMCs) and Continuous-Time Markov Chains (CTMCs) are used to represent systems with probabilistic transitions in discrete or continuous time, respectively [KNP07]. Additionally, Markov Decision Processes (MDPs) extend these models by incorporating nondeterministic choices [For+11], while Probabilistic Timed Automata (PTAs) combine real-time and probabilistic elements to capture systems with timing constraints [NPS13].

PRISM uses probabilistic temporal logic to specify properties for verification. These include: Probabilistic Computation Tree Logic (PCTL) for DTMCs and MDPs [KNP11]; and Continuous Stochastic Logic (CSL) for CTMCs [KNP02]. Users can define properties like "the probability of reaching a failure state within a certain time is less than 0.01" or "the expected cost of running the system until termination."

Models can be built either using PRISM's input language or importing models created with external tools. After constructing the model, PRISM performs model checking to verify whether the given properties hold, and it can compute performance measures like:

- Probabilities of events.
- Expected values (e.g., expected time to a failure).
- Long-run average costs or rewards.

PRISM also enables the analysis of reward-based properties, such as minimizing energy consumption or maximizing system reliability.

2.3.1 PRISM-Games

PRISM-games is an extension of the PRISM Model Checker, designed to handle the verification and strategy synthesis for multi-player stochastic games [Kwi+20]. It enables the modelling, verification, and analysis of systems where multiple entities (or "players") interact, potentially with conflicting goals, and where probabilistic outcomes influence the system's behaviour.

PRISM-games builds upon PRISM's capabilities by introducing game-theoretic aspects. It allows for the analysis of competitive and cooperative scenarios where players can adopt strategies to influence the outcome. The games are modelled using stochastic multi-player games (SMGs), where multiple players can interact in the system, or in other words: transitions between states depend not only on probabilistic outcomes but also on the strategies of multiple players. Such models enable the analysis of both zero-sum and non-zero-sum scenarios. Players can choose actions to influence the system's state transitions and the transitions can have probabilistic outcomes, reflecting uncertainty in the system.

PRISM-games also supports several types of models, including:

- Turn-based stochastic multi-player games (SMGs or TSGs): These extend the standard modelling of Markov Decision Processes (MDPs), as they can be viewed as a broader form of MDPs where each state is managed by an individual player.

- Concurrent stochastic multi-player games (CSGs): In this players make simultaneous decisions for transitions, with each player controlling specific modules and using actions uniquely tied to them.
- Turn-based probabilistic timed games (TPTGs): This extends PRISM-games by combining turn-based game features with Probabilistic Timed Automata (PTA) elements, specifying players with control actions and using clocks, guards, and invariants similar to PTAs.

PRISM-games extends the PRISM Model Checker to analyze multi-player stochastic games, enabling the synthesis of optimal strategies for players. These strategies can be *deterministic*, where a specific action is chosen in each state, or *randomized*, where actions are selected probabilistically. In multi-player scenarios, PRISM-games can compute Nash equilibria, a key concept in game theory where no player benefits by unilaterally deviating from their strategy.

PRISM-games enhances traditional probabilistic temporal logics, such as PCTL and CSL, to express goals specific to game settings. These logics allow users to specify objectives like ensuring a safe state is reached with a high probability, regardless of opponents' actions. Multi-objective verification further enables users to balance competing goals, such as maximizing rewards while minimizing risks, which is critical in safety-critical domains like autonomous systems.

PRISM-games was the first choice for implementing our reactive game approach.

2.3.2 Drawbacks of PRISM

The PRISM Model Checker is widely used for analyzing probabilistic systems, especially those with stochastic behaviours such as communication protocols, biological systems, and embedded devices. However, just like any other tool, PRISM and PRISM-games have limitations that can hinder their applicability.

One of the principal challenges of PRISM is the state space explosion problem, a frequent issue in model checking where the state space grows exponentially with the number of system components or variables. PRISM mitigates this with symbolic techniques, such as Binary Decision Diagrams (BDDs) and sparse matrix representations, but these approaches have limitations when applied to large systems [KNP11].

BI is the method of choice due to our finite horizon, as it enables optimal decision-making by considering future outcomes step-by-step from the end of the game back to the starting point. However, a major drawback of PRISM-games for our work is its inability to support Backward Induction effectively. Due to the size of the game tree,

Backward Induction can be computationally intense for large-scale games, especially when there are many decision points. While PRISM-games manages large state spaces symbolically, the lack of built-in Backward Induction further exacerbates computational challenges in extensive-form game models.

As described in section 2.2, BI typically requires recursive computation from terminal states, calculating each player's optimal response step-by-step back to the root of the game tree. PRISM-games, however, does not inherently support recursive solution methods tailored to BI, as its core algorithms are based on forward exploration of state spaces and policy synthesis, aiming to maximize or minimize objectives directly from initial states. As a result, PRISM-games cannot effectively represent the hierarchical structure of decisions needed for BI, thus making it less suitable in our approach.

2.4 MATLAB

MATLAB, created by MathWorks [MAT23], is a high-level programming environment widely used for data analysis, algorithm development, and numerical computation. MATLAB also offers an intuitive interface and powerful computational capabilities. MATLAB's matrix-based language facilitates the processing and visualization of complex datasets, while extensive toolboxes support specialized tasks, such as image processing, machine learning, and control systems.

Backward Induction(BI), introduced in section 2.2, is a fundamental method used in decision-making, game theory, and dynamic programming to solve problems that require optimizing decisions over multiple stages. MATLAB provides an ideal platform for implementing BI due to its robust computational capabilities, matrix-based computations, and extensive library of functions.

We can efficiently define and solve dynamic optimization problems using MATLAB by representing stages, states, and decisions as vectors or matrices. MATLAB's for loops and vectorized operations allow for concise implementation of recursive algorithms, minimizing computational overhead.

A key feature of MATLAB that supports BI is its ability to handle large-scale matrix operations, which is critical when solving problems with a large number of states or decision variables.

Overall, MATLAB's versatility and efficiency make it our preferred choice for implementing backward induction in our research project.

2.5 Cognitive Models

Cognitive models are theoretical frameworks designed to simulate and explain human cognition. They can represent processes such as perception, attention, decision-making, and memory, offering insights into how people process and respond to information. These models are grounded in cognitive science, psychology, and artificial intelligence, aiming to delineate how the mind processes information —ranging from perception and attention to decision-making and memory. They also allow researchers to predict human behaviour in various contexts, making them indispensable in fields like cognitive science, artificial intelligence, and human-computer interaction.

There are several types of cognitive models, each capturing different aspects of cognition. Our focus is on the various computational models that use mathematical notions to depict the various functions of human cognition. The cognitive models are classified into different types based on the chapter 'Cognitive Computing' by Gudivada [Gud16]:

- *Symbolic Models:* Rooted in classical cognitive science, symbolic models represent cognition through structured, rule-based systems that manipulate symbols in ways analogous to human reasoning. Examples include production systems like ACT-R (Adaptive Control of Thought-Rational) [And96; AL98], which simulates human thought processes through a series of condition-action rules that govern behavior; ICON FLUX [Niv07] which is used to simulate attention allocation in dynamic, complex systems; CASCaS (Cognitive Architecture for Safety-critical taskS) [WLB13] which simulate human decision-making and attention allocation in high-risk, high-demand tasks such as air traffic control or medical diagnostics; and PRIM (Perceptual Representation and Integration Model) [Taa13] focuses on how perceptual information is integrated and represented in the mind.

- *Connectionist Models (Neural Networks):* These models emphasize distributed, parallel processing similar to neural activity that occurs in the brain [FM14]. Rather than relying on predefined rules, connectionist models use networks of interconnected units (akin to neurons) to learn from data, adjusting the strength of connections between units based on experience. This approach aligns closely with the brain's plasticity and adaptability. Some of the examples include Convolutional Neural Networks (CNNs), Artificial neural networks (ANN), Deep Learning (DL), etc..

- *Hybrid Models:* Some cognitive models integrate both symbolic and connectionist elements, combining the rule-based precision of symbolic systems with the adapt-

ability of neural networks. This approach attempts to capture both high-level abstract reasoning and low-level perceptual or motor functions, offering a more comprehensive account of cognitive processes. Some of the hybrid models are: LIDA (Learning Intelligent Distribution Agent) [Fra+14], which models cognitive processes such as attention, perception, and memory using a symbolic structure, while connectionist mechanisms handle learning processes. LIDA incorporates attention and goal-directed behaviour with adaptive learning mechanisms to simulate intelligent behaviour in dynamic environments. IBM Watson [Fer+10], combines multiple technologies, including symbolic reasoning, machine learning, and natural language processing (NLP). Sigma (σ) [Ros13], is another hybrid model which integrates rule-based reasoning with probabilistic and distributed learning.

Cognitive models provide valuable frameworks for understanding and simulating human cognitive processes, ranging from perception and attention to decision-making and learning. These models, whether symbolic, connectionist, or hybrid, offer unique strengths in representing specific aspects of cognition, enabling researchers to analyze complex interactions and predict behaviour in various contexts. Another way of modelling humans is via their attention. One of the attention allocation models is the SEEV framework (Salience, Effort, Expectancy, Value). It provides a targeted approach to understanding how individuals prioritize and focus on information in dynamic environments. We will see more about this in the next section.

2.6 SEEV Attention Model

The SEEV attention model forms the basis of all our works. SEEV is an acronym for Salience, Effort, Expectancy, Value. The SEEV attention model [Wic+01], shown in fig. 2.1, is a comprehensive framework developed to explain how individuals allocate their attentional resources in dynamic and complex environments. Emerging from the domain of human factors and cognitive engineering, the SEEV model offers a structured understanding of attentional dynamics, which has been widely utilized to optimize system design and improve human performance across various applied fields.

The SEEV model identifies four key factors that interact to determine how attention is distributed: salience, effort, expectancy, and value. Each of these components plays a distinct role in shaping attentional behaviour.

- Salience (S): Salience refers to the extent to which a stimulus stands out from its surroundings due to its unique physical characteristics. Stimuli with higher

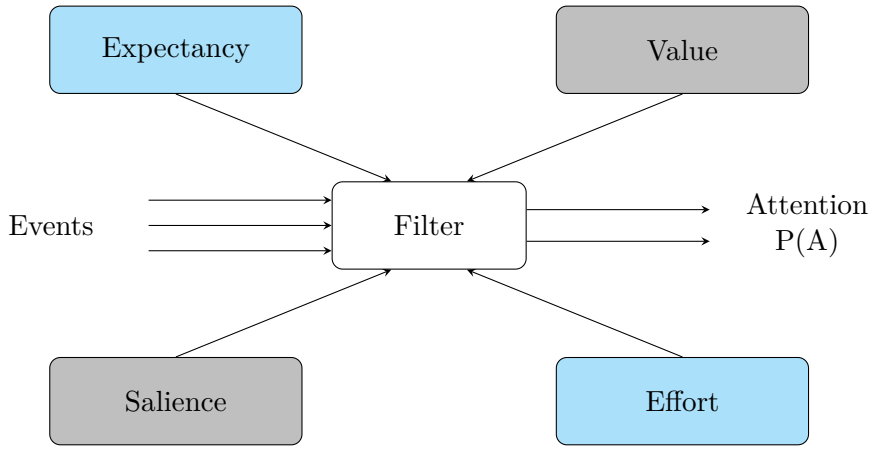


Figure 2.1: SEEV Model to determine attention by [Wic+01]

salience, such as bright colours, loud noises, or sudden movements, are more likely to attract attention automatically. This factor relies on the intrinsic properties of the environment to draw focus. For instance, in an aircraft cockpit, flashing warning lights or an alarm sound are designed to be salient enough to immediately capture the pilot's attention during critical situations [Wic15].

- Effort (Ef): Effort reflects the mental and physical resources required to shift attention from one location or stimulus to another. High effort costs can discourage frequent attention shifts, leading individuals to sustain focus on tasks or areas requiring minimal effort. For example, a driver may be less likely to glance at a distant road sign if the effort to shift attention away from the immediate roadway is perceived as too high [HWC06].

- Expectancy (Ex): Expectancy pertains to the likelihood of encountering a particular stimulus in a specific context based on prior knowledge or experience. This factor allows individuals to direct their attention more efficiently by predicting where and when relevant stimuli will appear. For example, a pilot scanning an instrument panel may prioritize certain gauges based on their relevance to the current flight phase, informed by training and experience [Wic+01].

- Value (V): Value represents the perceived importance or relevance of a stimulus to an individual's goals or objectives. High-value stimuli, deemed critical or beneficial, prompt greater allocation of attentional resources. For example, in a military operation, soldiers may prioritize attending to communication devices or mission-critical alerts over other distractions [Wor14]. Value is a flexible concept, shaped

by both immediate task demands and broader individual priorities, and is integral to decision-making in complex environments.

The *SEEV* model integrates these components into a quantitative framework for predicting the likelihood of attention (P(A)) being directed to a particular stimulus:

$$P(A) = S - Ef + Ex \cdot V \ . \tag{2.2}$$

The eq. (2.2), demonstrates the interplay of bottom-up factors (Salience and Effort) and top-down factors (Expectancy and Value) in influencing attention. Bottom-up factors are driven by the physical properties of the environment, while top-down factors are rooted in cognitive processes and an individual's learned expectations and experiences [Wic15].

The SEEV model has become a cornerstone in applied cognitive research, with significant contributions to fields such as:

- Aviation: In aviation, the SEEV model has been instrumental in studying how pilots allocate attention to various cockpit instruments and external cues. By understanding attentional priorities, designers can optimize instrument placement and display features to reduce errors and enhance situational awareness. For example, Wickens et al. [Wic+01] demonstrated how Salience and Value drive pilots' attention during critical flight operations, such as landing or responding to emergencies.

- Driving: The model has been extended to traffic environments to predict driver behaviour under varying conditions. Horrey et al. [HWC06] applied the SEEV model to explore how drivers allocate attention between the road, traffic signals, and in-vehicle devices, helping to design safer in-vehicle interfaces. Wortelen [Wor14] further refined this application to address driver distractions and improve attention management in Autonomous Vehicles.

- Military Operations: In military settings, where split-second decisions are critical, the SEEV model has been used to improve soldier performance by optimizing the design of command-and-control systems, by tailoring interfaces to align with natural attentional tendencies.

- Human-Machine Interface Design: Beyond specific domains, the SEEV model informs the design of user-friendly systems and interfaces. For instance, in medical devices or consumer electronics, aligning interface elements with attentional

principles ensures critical information is easily accessible, improving usability and reducing errors.

2.6.1 Modified SEEV Model

The SEEV (Salience, Effort, Expectancy, Value) model, introduced above, is widely used to predict attentional allocation based on environmental factors and task characteristics. It posits that attention allocation is a function of the Salience of visual stimuli, the Effort required to process those stimuli, the Expectancy of encountering relevant information, and the perceived Value of that information. In the modified SEEV model presented in this section (depicted in fig. 2.2), Salience and Effort are treated as constants, thereby emphasizing the dynamic interplay between Expectancy and Value in determining attentional focus.

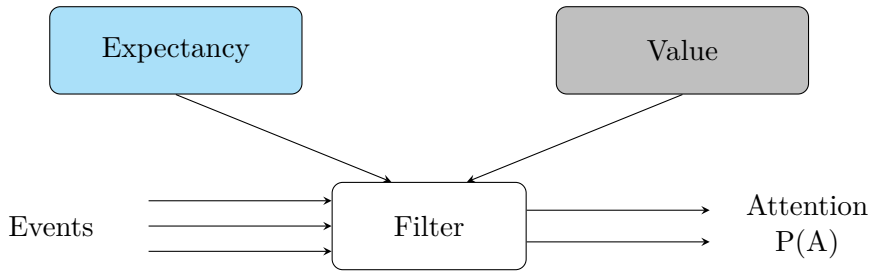


Figure 2.2: Modified SEEV Model

This modification stems from contexts where Salience and Effort are standardized or do not vary significantly between stimuli. For example, in experimental tasks designed to control visual and cognitive demands, all stimuli may be equally salient (e.g., standardized brightness, size, and contrast) and require uniform effort (e.g., the same physical or cognitive steps to engage). These conditions allow for isolating and analyzing the contributions of Expectancy and Value to attention allocation.

- *Salience as a Constant:* By keeping Salience uniform across stimuli, any differences in attentional allocation can be attributed to factors other than visual prominence.
- *Effort as a Constant:* Effort is held constant to eliminate variability due to task difficulty, ensuring that cognitive or physical demands do not disproportionately influence attention.
- *Expectancy and Value as Dynamic:* Expectancy and Value are treated as variable components, allowing us to explore how task goals, prior knowledge, or contextual cues influence attention allocation under controlled conditions.

$$P(A) = Ex \cdot V + c \ . \tag{2.3}$$

Equation (2.2) is modified to incorporate Salience and Effort as constants (c). This modified SEEV model is shown in eq. (2.3). Under this modified model, attention allocation becomes a direct function of Expectancy and Value. We will be using this modified SEEV model in all the future chapters of this dissertation.

3 Timing of an Atomic Explanation for a Single User

Contents

In this chapter, we introduce our framework for determining the optimal time for an atomic explanation production for a single user. The contents of this chapter are based on our work previously published in [Bai+22; BF24a].

3.1 Introduction

Autonomous systems, including vehicles, are becoming increasingly integral to daily life. However, as mentioned earlier, their complex and adaptive behaviours often surpass human understanding, presenting challenges in fostering trust and safety in human-machine interactions. Timely, effective explanations are crucial to addressing these challenges. While substantial research has explored the content of explanations [Koo+16; Wie+19], there is a limited understanding of the optimal timing for explanation delivery. This section focuses on leveraging the SEEV (Salience, Effort, Expectancy, Value) attention model, introduced in section 2.6, within a reactive game framework to optimize explanation timing in Autonomous Vehicles (AVs). By modelling atomic explanation delivery as a Markov Decision Game using Markov Decision Process (MDP), we synthesize strategies that minimize the cognitive workload of the user.

This chapter is divided as follows: In section 3.2, we introduce a running example for this chapter, which shows the intricate relationship between timing and explanation. In section 3.3, we provide an overview of the fundamental concept of SEEV, followed by

the implementation in a reactive game. The results of the reactive game are discussed in section 3.4, and we finally conclude this chapter by summarizing the key contributions in section 3.5.

3.2 Example Scenario

To illustrate how the timing of an explanation affects human attention, consider the following scenario from Bairy et al. [Bai+22], depicted in fig. 3.1:

Example 3.2.1. An Autonomous Vehicle v is approaching an intersection where it plans to make a left turn. The traffic light is green, signalling that it is safe to proceed without interruption. However, the vehicle unexpectedly comes to a stop. The reason for the stop is that an emergency vehicle is approaching, and v is yielding to it. v intends to explain to its passengers that the stop is due to yielding to an emergency vehicle, but it must decide when and whether to offer this explanation.

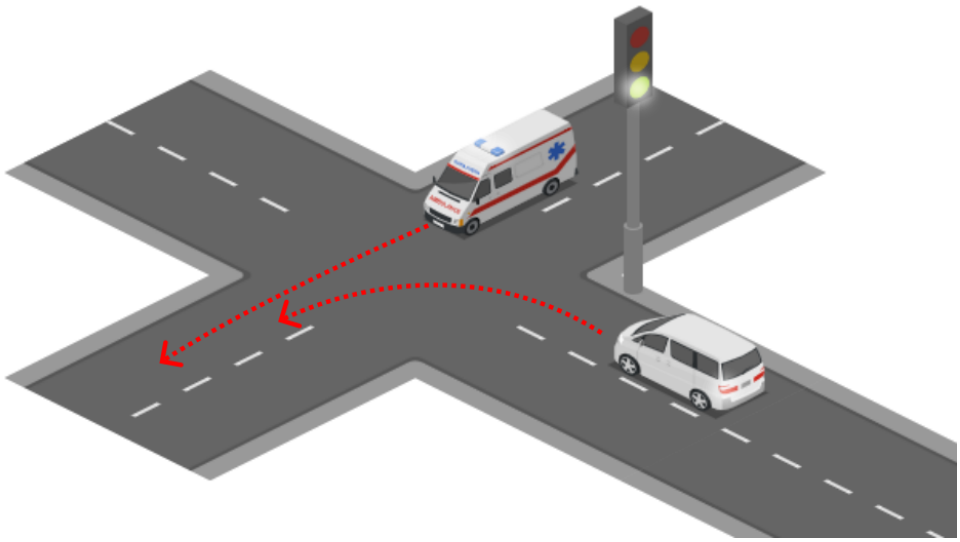


Figure 3.1: AV waiting example: AV waiting at a green traffic light

The timing of this explanation is crucial. If the vehicle delivers the explanation too early ——for example, before passengers have had a chance to recognize that they are approaching an intersection and that a left turn is imminent—— the explanation might be perceived as irrelevant or premature. At this stage, the passengers have not yet encountered any confusion, so the information might be dismissed, and the cognitive effort involved in processing it could be wasted.

On the other hand, if the explanation is provided too late ——or if it is not provided at all—— the passengers may begin to wonder why the vehicle stopped unexpectedly at a green light. As a result, they will likely engage in their own search for an explanation, by scanning the environment and considering possible reasons for the stop. This increased mental effort can lead to a higher cognitive workload, as passengers try to make sense of the situation. The delay in receiving the explanation forces them to focus their attention on their immediate surroundings, which could increase their anxiety or confusion as they try to find a solution independently.

Providing an explanation too soon may overwhelm the passengers with unnecessary information when they are not yet focused on the situation at hand. And providing an explanation too late would result in them starting their own attention strategy. Therefore, the successful transfer of understanding between the vehicle and the passengers depends not only on the content of the explanation but also on the precise timing of when it is delivered.

3.3 SEEV Model in a Reactive Decision Game

As sketched in section 3.1, to optimize the timing of explanations, this study models the decision-making process as a *1.5-player game*——a Markov Decision Game using Markov Decision Process (MDP) [How60] involving a strategic player (the AV's explanation mechanism) and a random player (the human's attention modelled by SEEV). The details of the SEEV modelled can be found in section 2.6.

In this chapter, we are going to use the modified SEEV model from section 2.6.1 as we are focusing on the example 3.2.1. We will be considering only the top-down factors: Expectancy (Ex) and Value (V). The Salience (S) and Effort (Ef) are considered to be constant in this situation as in example 3.2.1, the area of interest in fixed and hence the Effort factor remains constant. Given the short time span of the scenario, we can also approximate the salience as being constant throughout the scenario. Thus we use the modified SEEV formula, given in eq. (2.3), in this chapter.

In the reactive game implemented here, the strategic player determines when to provide an explanation, aiming to minimize cognitive workload, while the random player's behaviour, governed by the SEEV model, influences the dynamics of attention allocation and workload. The SEEV model's stochastic nature enables the random player to probabilistically determine attention levels based on temporal and contextual factors. The strategic player must decide between three states:

1. *No explanation (no_expl):* In this state, the strategic player (AV) withholds an explanation, assuming it is not immediately required or beneficial. This action may occur when:
   - The AV predicts that providing an explanation at this moment would induce unnecessary cognitive workload because the human's attention is not yet focused on the scenario.
   - The scenario's urgency or complexity does not yet justify interrupting the human's cognitive process.

   However, refraining from delivering an explanation is not without cost. If the human begins actively scanning the environment to deduce the reason behind the AV's behaviour (e.g., stopping unexpectedly), this can significantly increase cognitive workload. The decision to stay in this state must therefore account for the risks of workload escalation due to the absence of timely information.

2. *Explanation (expl):* In this state, the AV provides an explanation to the human. This action introduces cognitive workload because the occupant must process and comprehend the explanation. The workload induced depends on factors such as the timing, complexity, and perceived relevance of the explanation.
   - If delivered at the right moment, the explanation preempts the human's need to search for environmental cues, reducing the overall workload.
   - If delivered too early, the explanation may be disregarded, leading to wasted cognitive effort and a potential need for repetition.
   - If delivered too late, the human may already have engaged in an attention-intensive search for answers, rendering the explanation less effective or redundant.

3. *Explanation unnecessary (no_expl_needed):* This state represents situations where an explanation is deemed redundant or irrelevant. Scenarios that justify this state include:
   - The human has already observed sufficient environmental cues to understand the AV's actions without additional clarification. For instance, the sight of an emergency vehicle approaching may make the reason for the AV's halt self-evident.

- The situation resolves itself naturally, eliminating the need for an explanation. For example, a delayed start at a green light might become understandable as traffic begins moving.
- The context or timing indicates that providing an explanation would not meaningfully enhance the human's understanding or reduce their workload.

Entering this state is advantageous when the AV can accurately predict that an explanation would neither prevent cognitive effort nor improve situational clarity.

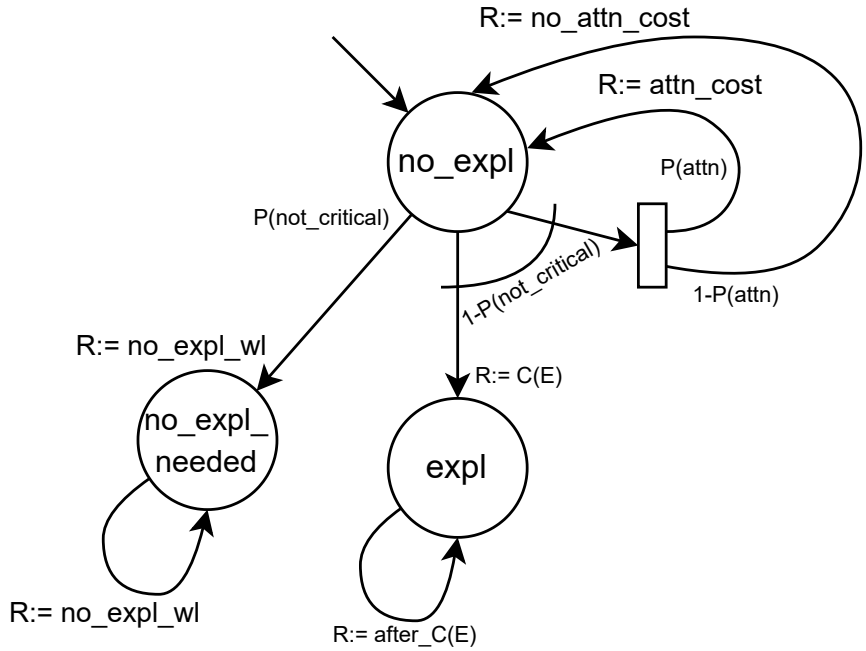


Figure 3.2: MDP representing the options for strategic player; to be combined be a product construction with the SEEV model in order to obtain the actual game graph

The MDP of the options for strategic player is illustrated in fig. 3.2. This figure visualizes the state transitions of the AV's explanation mechanism. It outlines the possible actions the strategic player can take at each timestep (we consider one timestep to be equivalent to 1 s), transitioning between the states (discussed above) of providing, withholding, or identifying unnecessary explanations. Rewards or costs are associated with these transitions, which are calculated based on attention probabilities derived from SEEV and scenario-specific conditions.

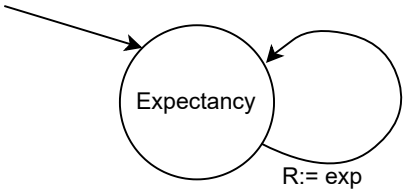


Figure 3.3: Expectancy of the random player

The Value (V) of the SEEV model may vary when transitioning to a differently structured environment (e.g., from urban driving to an expressway). However, in this study, we treat V as a constant because the game is restricted to the example scenario in example 3.2.1 which involves only a single area of interest (the intersection). Hence, as shown in fig. 3.3, the random player's Expectancy (Ex) accumulates over time. Ex grows linearly.This progression reflects the increasing cognitive demand as the human user anticipates an explanation. By examining the expectancy curve, the AV can strategically time its explanations to preempt the peak cognitive workload, thereby optimizing the interaction. In the next section, we discuss more about how the rewards are calculated.

3.4 Game Results

The primary objectives of this reactive game model were twofold: to determine the optimal timing for delivering explanations to achieve the lowest cognitive workload on the human and to evaluate the minimum expected workload across all possible presentation strategies. These evaluations required assigning rewards (which mirror workload) to the transitions between states, as outlined in table 3.1.

S	S'	Probability	R
no_expl	no_expl	P(not_critical) ·P(A)	0.4
no_expl	no_expl	P(not_critical) ·(1 - P(A))	0.2
no_expl	expl	P(not_critical)	0.3
no_expl	no_expl_needed	P(not_critical)	0.0
expl	expl	1	0.1
no_expl_needed	no_expl_needed	1	0.0

Table 3.1: MDP rewards

The reactive game model is a finite horizon model, and hence for this model, the minimum workload at any given time was calculated using backward Bellman induction BI, introduced in section 2.2. This recursive process considered the cost of presenting

or withholding explanations at each timestep (1 s). The minimum workload (min_wl) depends on several factors, including the probability of a critical scenario not occurring ($P(not\_critical)$) and the cost associated with providing or not providing explanations. The formula used for this computation is given by eq. (3.1):

$$min\_wl_n^k = P(not\_critical) \cdot no\_expl\_wl + (1 - P(not\_critical)) \cdot expl\_wl_n^k \quad (3.1)$$

Here, no_expl_wl is a constant that is employed when no explanation is required. $expl\_wl_n^k$ represents the cost associated with explanation delivery at time n and k represents the total duration of the scenario. It factors in whether attention is being paid or not and incorporates a recursive reduction of the temporal horizon. This value is calculated as shown in eq. (3.2):

$$expl\_wl_n^k = min \begin{cases} C(E) + (k-n) \cdot after\_C(E), \\ P(A)_n \cdot (min\_wl_0^{k-n} + attn\_cost) \\ +(1 - P(A)_n) \cdot (min\_wl_{n+1}^k + no\_attn\_cost) \end{cases} \quad (3.2)$$

Here $C(E)$ represents the cost of providing the explanation, $after\_C(E)$ is the ongoing cognitive cost that is incurred after the explanation is provided, $P(A)$ depicts the probability of attention that is being paid, $attn\_cost$ is the workload cost associated with the human actively attending to a scenario (by scanning the environment), and no_attn_cost represents the workload incurred when the human does not actively pay attention.

Based on these rewards, the optimal time to provide an explanation (t_expl) and the corresponding minimum workload (min_wl) for various times until the scenario occurs (t_max), are presented in the table 3.2. Here t_expl is the time to provide the explanation from the current moment.

The reactive game was built using MATLAB [MAT22]. Simulation results, shown in table 3.2, reveal that optimal explanation timing is highly scenario-dependent. For scenarios lasting less than 2 s, no workload reduction can be achieved through early explanations. However, as the time horizon increases, explanation timing becomes critical.

For short scenarios, a single explanation close to the event is optimal. For instance, when $t\_max = 5$ s, the ideal explanation time is 2 s before the event. When extending scenario durations further, we make the interesting observation that up to scenario duration $t\_max = 15$ s, the optimal explanation time is 3 s before the scenario occurs. Interestingly, as the scenario duration extends, providing multiple explanations becomes

t_max (s)	t_expl (s)	min_wl
2	2	0.300
3	2	0.400
4	2	0.500
5	2	0.500
6	3	0.600
7	4	0.600
8	5	0.600
9	6	0.600
10	7	0.600
11	8	0.600
12	9	0.600
13	10	0.600
14	11	0.600
15	12	0.600
16	2, 13	0.600

Table 3.2: Optimal explanation time (t_expl) generating minimum expected workload (min_wl)

necessary. For scenarios lasting 16 s or more, the optimal strategy involves delivering an initial explanation early (e.g., at 2 s) and another closer to the event (e.g., 3 s before occurrence). This will be explored in more detail in the next chapter.

This behaviour underscores the complexity of explanation timing. Contrary to intuition, the earliest possible explanation is not always the most effective. Instead, the optimal timing aligns with specific scenario dynamics and attention probabilities.

Regarding computational feasibility, the backward induction process proved efficient for real-time application in shorter scenarios ($t\_max \leq 20$ seconds), with computation times remaining under 1 s. However, for longer scenarios, the recursive nature of the algorithm leads to exponentially increasing computation times. Consequently, such cases may require offline computation or more efficient execution platforms, rather than MATLAB, for practical implementation.

3.5 Chapter Summary

In this chapter, we introduced our reactive game framework integrated with the SEEV attention model to optimize explanation timing in AVs. Results demonstrate the nuanced interplay between explanation timing and cognitive workload, emphasizing that neither immediate nor delayed explanations uniformly minimize workload. The results

presented in the previous section are based on the cost/reward values postulated for the different transitions shown in table 3.1. These costs currently are just educated guesses serving the purpose of demonstration of the technology, yet lack empirical psychological grounding. In the next chapter, we explore the extension of the framework to multi-step explanations.

4 Timings of Multi-Step Explanations for a Single User

Contents

In the results presented in chapter 3, we observed that our model tends to propose multi-step explanations for scenarios lasting 16 s or longer. However, a key question arises: in situations where complete information is unavailable at the start, should we wait to provide a comprehensive explanation, or should we offer partial information upfront? This chapter explores this dilemma. To address it, we extend the framework introduced in chapter 3 to include multi-step explanations. Some parts of this chapter have been published in [BF23].

4.1 Introduction

Krull and Anderson[KA97] and El-Assady et al. [El-+19] suggest that explanation is not necessarily a single step process. We had previously proposed that explanation is a dynamic process of belief updates and hence would need multi-step explanations [Bai+22]. In the last decade, there has been an increase in research related to dialogue-based explanations [DO22; Xu22; Xu+23; Min+24]. However, before tackling the complex task of implementing dialogue-based explanations, it is helpful to first determine the appropriate timing for presenting multi-step explanations. This chapter explores when it is most suitable to offer such explanations.

This chapter is divided as follows: In section 4.2, we introduce a running example for this chapter, which exemplifies the need for multi-step explanations. In section 4.3, we

provide an overview of the fundamental concept of SEEV, followed by its implementation in a reactive game. The results of the reactive game are discussed in section 4.4, and we finally conclude this chapter by summarizing the key contributions in section 4.5.

4.2 Example Scenario

Explanations often consist of multiple pieces of information, and their timing can vary depending on how much information is available at a specific moment. In scenarios where information is incomplete or evolving, deciding when and how to communicate details becomes crucial, as it directly impacts how effectively the explanation serves its purpose. To illustrate this, consider the following example which is also depicted in the fig. 4.1:

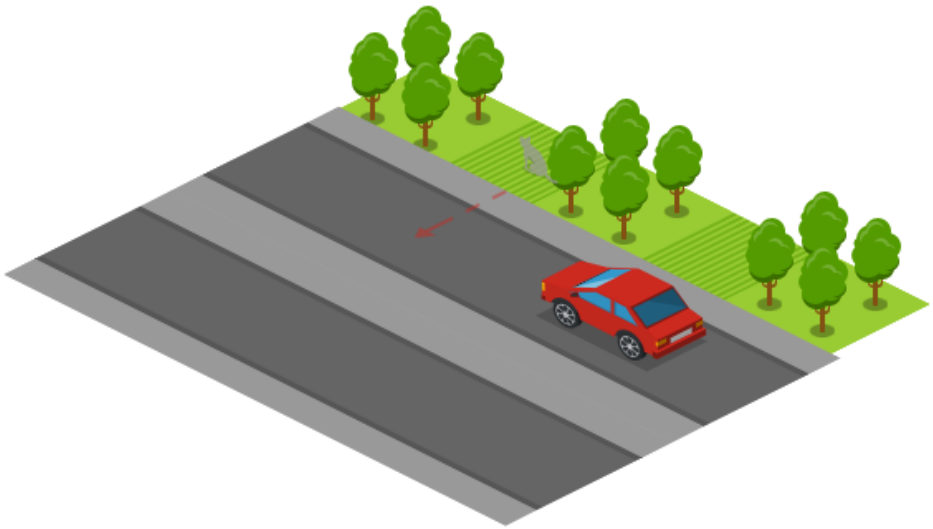


Figure 4.1: Potential hazard example: potential hazard on the road

Example 4.2.1. An Autonomous Vehicle v is navigating a road when it detects a potential hazard ahead. Based on its sensors and perception algorithms, v recognizes that something is obstructing its path and initiates a deceleration manoeuvre to maintain safety. However, at this stage, the system has only partial information about the nature of the hazard —it could be a cyclist, an animal or an inanimate object, but v 's perception components have not yet determined the specific classification.

This scenario highlights a situation where the AV has only partial information about the nature of the potential hazard at the current time. Based on these incomplete data, the AV could choose to provide an explanation to the human passenger, detailing

the detected hazard and the uncertainty surrounding its classification. While such an explanation may reduce the human's cognitive workload by offering transparency, it could also increase their cognitive burden. This increase might occur if the human passenger begins to actively engage in their own attention and reasoning strategies to address the incomplete information.

Alternatively, the AV might delay providing an explanation until it has fully identified the hazard. While this approach ensures that the explanation is more comprehensive and definitive, it introduces a different challenge: the delay could cause uncertainty or anxiety for the human, especially if they have already started forming their own attention strategies. Moreover, if the complete information becomes available too late, the delayed explanation may fail to mitigate cognitive workload effectively or support the human's situational understanding.

4.3 Reactive Game using SEEV Model

This chapter, as outlined in section 4.1, models a decision-making process as a *1.5-player game* to optimise the timings of multi-step explanations. Specifically, it is modelled as a Markov Decision Process (MDP) [How60], involving two key players: a strategic player, represented by the AV's explanation mechanism, and a random player, capturing human attention as modelled by the SEEV framework.

In this chapter, we employ the modified SEEV model introduced in section 2.6.1 to focus on the scenario described in example 4.2.1. We apply the modified SEEV formula presented in eq. (2.3), as we consider Effort and Salience to be constants.

For ease of implementation, we consider the explanation to comprise two distinct parts of information. In the reactive SEEV game, the process begins n seconds before the onset of the scenario and concludes upon its completion. The first part of the information becomes available n seconds prior to the scenario's occurrence, while the second part becomes accessible halfway through this period, at $n/2$ seconds before the scenario begins.

Building on the SEEV model, we expand the reactive game framework from section 3.3. The reactive game considers two key factors: the asynchronous availability of partial and complete information; and the influence of explanation timing on human cognitive workload, incorporating costs associated with receiving explanations and pursuing independent attention strategies.

In the reactive game, the strategic player (AV) has four primary actions at each time step:

1. *No explanation (no_expl):* The AV refrains from providing any explanation, preserving cognitive resources but potentially increasing the passenger's mental workload as they attempt to interpret the situation independently.

2. *Partial explanation (expl*1*):* The AV provides available partial information, giving some clarity but potentially leaving the passenger with unanswered questions. We assume that the AV has access to this partial information at the start of the scenario.

3. *Complete explanation (expl*2*):* The AV delivers a full explanation once all necessary information becomes available, ensuring comprehensive understanding but potentially inducing a higher workload due to delayed timing. The full explanation becomes available after half the scenario has been completed.

4. *No explanation needed (no_expl_needed):* The AV identifies that no explanation is required, as the situation is either resolved or does not warrant clarification. This action minimizes unnecessary cognitive load.

The MDP of the strategic player's decision-making process is shown in fig. 4.2. States represent the AV's options: withholding explanations (*no_expl*), providing partial explanations (*expl*1), delivering complete explanations (*expl*2), or determining no explanation is needed (*no_expl_needed*). Transitions between states are governed by workload costs and rewards. Some of the transitions are shown below:

- Transitioning from *no_expl* to *expl*1 reflects the workload cost of offering partial information.

- Moving from *expl*1 to *no_expl* and then to *expl*2 involves updating the explanation with complete information, which may yield rewards for enhancing clarity.

- Selecting *no_expl_needed* terminates the need for further action, minimizing unnecessary cognitive load

The random player —human attention— reacts dynamically based on the SEEV model. Attention probabilities (P(A)) fluctuate according to the buildup of Expectancy, shown in fig. 4.3, and the perceived value of information, influencing the strategic player's decisions. The probability of attention (P(A)) and the probability of no critical scenario ($P(not\_critical)$) help in determining the reward values for the strategic player. The details about the reward structure is discussed more in the next section.

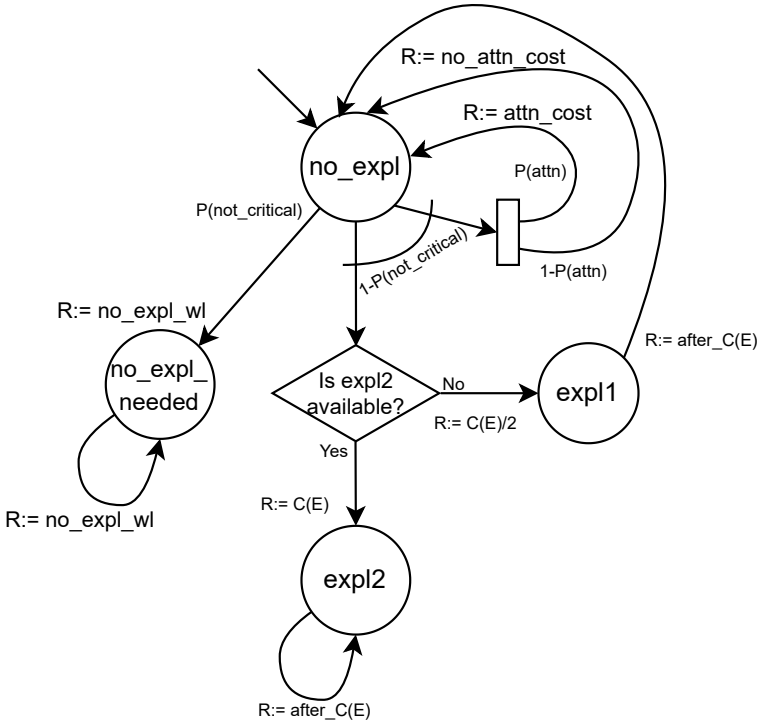


Figure 4.2: MDP representing the options for strategic player with two explanations; to be combined be a product construction with the SEEV model in order to obtain the actual game graph

The game is implemented in MATLAB [MAT22] to simulate the example scenario from section 4.2 for varying lengths of the time horizon and calculate optimal strategies. The implementation involves Backward Bellman Recursion and Reward Optimisation.

- *Backward Bellman Recursion (BI):* To compute the minimum workload (min_wl) at each state and time step, BI evaluates the costs of immediate versus delayed explanations.
- *Reward Optimisation:* The system identifies the optimal strategy by maximizing rewards tied to workload reduction and passenger comprehension.

By analyzing the reward structure and workload costs, the framework identifies the optimal timing for explanations, which adapts based on scenario dynamics and the time horizon until the event.

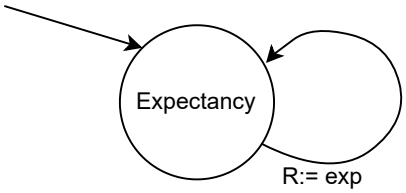


Figure 4.3: Expectancy of the random player

4.4 Game Results

Our goals for this reactive SEEV game were to identify whether, given that the AV determines information contributing to the explanation asynchronously in two parts, the explanation should indeed be provided in two parts and, if so, to determine the optimal timing for each part to minimize the cognitive workload of the human. To achieve this, a cost structure detailing cognitive workload costs was assigned to various state transitions for computation. This cost structure, as outlined in table 4.1, enabled the calculation of optimal explanation strategies by associating probabilities and rewards with different transitions.

S	S'	Probability	R
no_expl	no_expl	(1 - P(not_critical)) ·P(A)	0.4
no_expl	no_expl	(1 - P(not_critical)) ·(1 - P(A))	0.2
no_expl	is expl2 available?	(1 - P(not_critical))	0.3
no_expl	no_expl_needed	P(not_critical)	0.0
expl1	no_expl	1	0.05
expl2	expl2	1	0.1
no_expl_needed	no_expl_needed	1	0.0

Table 4.1: MDP rewards

The evaluation utilized the scenario described in the example section, focusing on varying lengths of the scenario to simulate different time horizons. By analyzing state transitions, rewards, and costs, the experiments quantified the impact of explanation timing on cognitive workload and validated the proposed framework using backward Bellman induction and dynamic programming methods. The minimum workload (min_wl) is given by eq. (4.1). The probability of a non-critical scenario ($P(not\_critical)$) accounts for situations where the user may have already assessed the environment, or where the critical situation was resolved through the temporal progression of events. A constant workload value associated with $P(not\_critical)$ is represented by no_expl_wl. In cases

where an explanation is necessary, the workload ($expl\_wl_n^k$) is calculated through backward induction, with k indicating the time of the scenario and n representing the current time. This calculation is described in Eq. 4.2.

$$min\_wl_n^k = P(not\_critical) \cdot no\_expl\_wl + (1 - P(not\_critical)) \cdot expl\_wl_n^k \quad (4.1)$$

If the total scenario time is k, $expl\_wl_n^k$ represents the minimum between the costs incurred by providing an explanation now and the costs incurred by not providing one. This reflects the strategic decision aimed at minimizing the expected workload. Waiting for an explanation incurs a cost ($waiting\_cost$), which is determined by the Probability of Attention (P(A)) derived from the SEEV model in eq. (2.3). This cost is calculated using backward Bellman recursion.

$$expl\_wl_n^k = min \begin{cases} expl\_cost, \\ P(A)_n \cdot (min\_wl_0^{k-n} + attn\_cost) \\ +(1 - P(A)_n) \cdot (min\_wl_{n+1}^k + no\_attn\_cost) \end{cases} \quad (4.2)$$

When the human is paying attention, $waiting\_cost$ is the cost associated with following an attention direction ($attn\_cost$), combined with the minimum workload over the reduced horizon ($k - n$). When the human is not paying attention, $waiting\_cost$ represents the cost of not paying attention (no_attn_cost), along with the minimum workload obtained through backward recursion.

$$expl\_cost = \begin{cases} C(E) + (k - n) \cdot after\_C(E) & expl2 \ is \ available \\ 0.5 \cdot C(E) + (k - n) \cdot after\_C(E) & otherwise \end{cases} \quad (4.3)$$

Providing an explanation incurs a cost (expl_cost) as shown in Eq. 4.3. This cost is the sum of the cost of receiving an explanation ($C(E)$) and a constant cost that arises after an explanation is provided ($after\_C(E)$). In our model, two types of explanations are available: one with partial information at the start and another with complete information provided later. The cost/reward of the explanation depends on which type is given. If the first (partial) explanation is provided, only half of the reward, $C(E)$, is granted. Table 4.2 summarizes the minimum workload (min_wl) and optimal times (t_expl1/t_expl2) for providing explanations at different times until the scenario occurs (t_max), based on the reward structure mentioned above.

t_max (s)	t_expl1 (s)	min_wl for expl1	t_expl2 (s)	min_wl for expl2	CPU time (s)
2	-	-	2	0.300	0.0100
3	-	-	2	0.400	0.0100
4	-	-	2	0.500	0.0200
5	-	-	2	0.300	0.0200
6	-	-	3	0.500	0.0300
7	-	-	4	0.600	0.0400
8	-	-	5	0.600	0.0600
9	-	-	6	0.600	0.0700
10	2	0.550	7	0.600	0.0800
11	2	0.600	8	0.600	0.1800
12	2	0.650	9	0.600	0.2700
13	2	0.700	10	0.600	0.3500
14	2	0.750	11	0.600	0.6200
15	2	0.800	12	0.600	0.8600

Table 4.2: Optimal explanation times for 2 explanations based on minimum workload

Table 4.2 presents the results of optimizing the timing of explanations for horizons ranging from 2 s to 15 s. For events that occur within a second or are already happening, the model suggests that an early explanation does not lead to a reduction in workload. However, as the time to the event increases, the timing of the explanation becomes crucial. It is neither optimal to explain as early as possible nor as late as possible; rather, the timing of the explanation follows a piecewise affine function based on the duration of the scenario. Contrary to intuition, it is not always best to provide an explanation immediately. There exists a specific point during the scenario when it is most beneficial to provide the explanation. For horizons up to 9 s, despite part of the explanation being available earlier, it is best to provide the explanation closer to the event —specifically, 3 s before the scenario occurs, when all the information is available. However, for horizons of 10 s or more, the approach shifts: a partial explanation is useful when given at 2 s, followed by a complete explanation 3 s before the scenario occurs.

Table 4.2 also reports the computation runtimes for scenarios with varying t_max values. Due to the backward induction process, the computation time increases exponentially with the horizon length. Despite this, even using MATLAB, which is not the most efficient execution platform, the computation time remains under 1 s for t_max values up to 15 s. This ensures that the algorithm can be executed in real-time. For larger t_max values, however, the algorithm may need to be run offline or on a more efficient execution platform.

4.5 Chapter Summary

In this chapter, we presented a novel approach to optimising explanation timing for multi-step explanations in AVs using the SEEV model integrated into a reactive game framework. Experimental results emphasize the significance of adapting explanation timing based on the time horizon and information availability. Both short ($\leq 9\,\mathrm{s}$) and long horizons ($> 9\,\mathrm{s}$) showcased unique strategies to minimize cognitive workload while enhancing passenger comprehension.

5 An Atomic Explanation for Multiple Users

Contents

Until now, we have focused on determining the optimal timing for delivering atomic or multi-step explanations to a single user. However, in real-world scenarios, it is common for multiple individuals to be present during a drive. This necessitates extending our explanation timing framework to accommodate multiple users. In this chapter, we explore how to adapt the framework for presenting atomic explanations to two users. The content of this chapter is based on our previously published work in [BF25].

5.1 Introduction

The SAE levels of driving automation (0-5) highlight the progressive transition from human control to full automation [Int21]. At higher levels (4-5), the timing and delivery of explanations become critical to maintaining passenger confidence, especially in complex or unexpected scenarios. Poorly timed explanations —either too frequent or delayed— can undermine user trust and comprehension. As interactive systems increasingly involve multiple users with varying attention patterns and cognitive states, the need for precise explanation timing becomes essential. In this chapter, we build on the SEEV (Salience, Effort, Expectancy, Value) attention model, introduced in section 2.6, to optimize explanation timing for multiple users.

Previous chapters have explored explanation timing for single users using game-theoretic approaches. Shen et al. [She+20] emphasized the importance of delivering explanations in critical situations, with pre-action explanations found to enhance trust [Du+19;

RTC18]. Körber et al. [KPB18] noted that even delayed explanations can improve user understanding, though pre-action explanations are generally more effective. Despite these insights, multi-user scenarios remain underexplored.

This chapter extends the prior model (chapter 3) to a multi-user context, utilizing a Markov Decision Process (MDP) framework to determine optimal explanation timing. The approach employs backward Bellman induction to calculate strategies that minimize cognitive strain, offering practical solutions for multi-user systems in AVs and beyond. In section 5.2, we provide an overview of the fundamental concept of SEEV for two users, followed by its implementation in a reactive game. The results of the reactive game are discussed in section 5.3, and we finally conclude this chapter by summarizing the key contributions in section 5.4.

5.2 Reactive Game Model

The timing optimization of explanation delivery for multiple users in an Autonomous Vehicle (AV) setting is achieved using a reactive game. For simplicity, this chapter considers two users. The game involves the AV and mental workload models for these users, acknowledging the link between mental workload and attention. Kantowitz [Kan00] asserts that even simple attention models can predict mental workload effectively. In this work, we use the SEEV attention model [Wic+01], detailed in section 2.6, to measure user attention.

5.2.1 SEEV Model for Two Users

The general overview of the SEEV model can be found in section 2.6. Here, we have made some slight modifications to the different attributes of the SEEV model; these adjustments have been proposed by Wortelen [Wor14]. Below is an explanation of each attribute, which has been tailored for the current work:

1. **Salience (S):** This refers to the degree to which new information of a specific type will attract the attention of an individual if it becomes available. In other words, it measures how noticeable or attention-grabbing the information is.

2. **Effort (Ef):** Effort denotes the amount of physical or cognitive energy required by the individual to process the new information. This could include the mental effort needed to comprehend or the physical effort required to interact with the information.

3. **Expectancy (Ex):** Expectancy represents the predicted frequency with which new information will appear. It acts as a dynamic variable, offering an estimate of how much time will pass before updated information is received.

4. **Value (V):** Value signifies the anticipated benefit or gain that the individual expects to derive from the information. It represents the perceived usefulness of the information and the extent to which the user believes it will provide a return, whether in the form of knowledge, rewards, or other positive outcomes.

In this chapter, we apply the SEEV model to represent two distinct user profiles. We achieve this by adjusting the expectancy attribute within the SEEV framework. These adjustments lead to the development of different emotional states in users, such as anxious, calm, or bored, which allows for a more detailed and varied analysis of user experiences.

In our model, we begin by randomly generating initial Expectancy values for each user. This is meant to model the history preceding the scenario. These initial values are constrained to be less than 0.5. To determine the final expectancy values, we generate a uniformly distributed random number r within a specified range shown in 5.1.

$$r = \mathcal{U}(a, b) \tag{5.1}$$

Using MATLAB, we generate the final expectancy values as follows:

$$r = a + (b - a) * rand(N, 1) \ . \tag{5.2}$$

Here, r represents the final expectancy value, a and b define the lower and upper bounds of the interval, respectively, and *rand(N,1)* generates a column vector of N random numbers between 0 and 1. The parameter N represents the number of random values to be generated, and in our model, N is set to 1 since each user will receive a single final expectancy value. The resulting random number will be scaled to fall between a and b, which allows for a controlled variation of expectancy values.

Once we have determined both the initial and final Expectancy values for each user, as well as the corresponding initial and final time values, we use the *fit function* in MATLAB to model the relationship between these variables using an exponential curve. Equation (5.3) describes the fit function.

$$f = fit(x, y, `exp1') \tag{5.3}$$

In this equation, x represents the array of initial and final Expectancy values, while y represents the array of corresponding time values. The initial time value is set to 1, while the final time value corresponds to the moment when the scenario occurs. This final time value can be adjusted to generate different Expectancy profiles. The argument *'exp1'* specifies the type of fit function used, which is a single-term exponential curve, as shown in eq. (5.4).

$$f(x) = a * exp(b * x) \tag{5.4}$$

This equation describes an exponential function where a and b are parameters determined by the fitting process, and x represents the input values (Expectancy). The function models how the Expectancy changes over time, capturing the growth or decay of Expectancy based on the parameters a and b. This fitting approach allows us to create a dynamic model that simulates how Expectancy evolves throughout the scenario.

5.2.2 Model Implementation

In safety-critical scenarios, the timing of explanations is vital to balance user comprehension with avoiding unnecessary distraction or confusion. Consider the following example (example 3.2.1) adapted from [BF24a] where, at an intersection, an Autonomous Vehicle (v) intends to make a left turn but comes to a stop, despite the green light permitting an uninterrupted turn. The vehicle has detected an approaching emergency vehicle and stops to yield the right of way. Although v can explain its actions to occupants, it must carefully decide when and how to provide this explanation. This is the same example from chapter 3. In chapter 3, we assumed that there was only one user in the v. Here, we consider two users to be present in v.

In this scenario, an explanation given too early —before users recognize the intersection or the planned left turn— may be ignored, as it does not align with their current understanding of the situation. This misalignment wastes cognitive effort. Conversely, delivering the explanation too late, or not at all, may lead users to engage their own attention strategies, such as scanning the environment to infer why the vehicle stopped. This increases cognitive workload unnecessarily.

Minimizing cognitive load thus relies on delivering the explanation at the optimal moment, ensuring it is both timely and relevant. This prevents unnecessary mental effort while keeping users informed and reassured.

For the example 3.2.1, we can consider Salience and Effort to me constant. Hence we can use the modified SEEV eq. (2.3) from section 2.6.1.

This simplified formula has been integrated into a Markov Decision Process (MDP) game, where decisions centre on the timing of explanation delivery. Using MATLAB [MAT23], we calculated reactive strategies for explanation timing. The SEEV-based game begins k seconds before the scenario starts and concludes when the scenario ends.

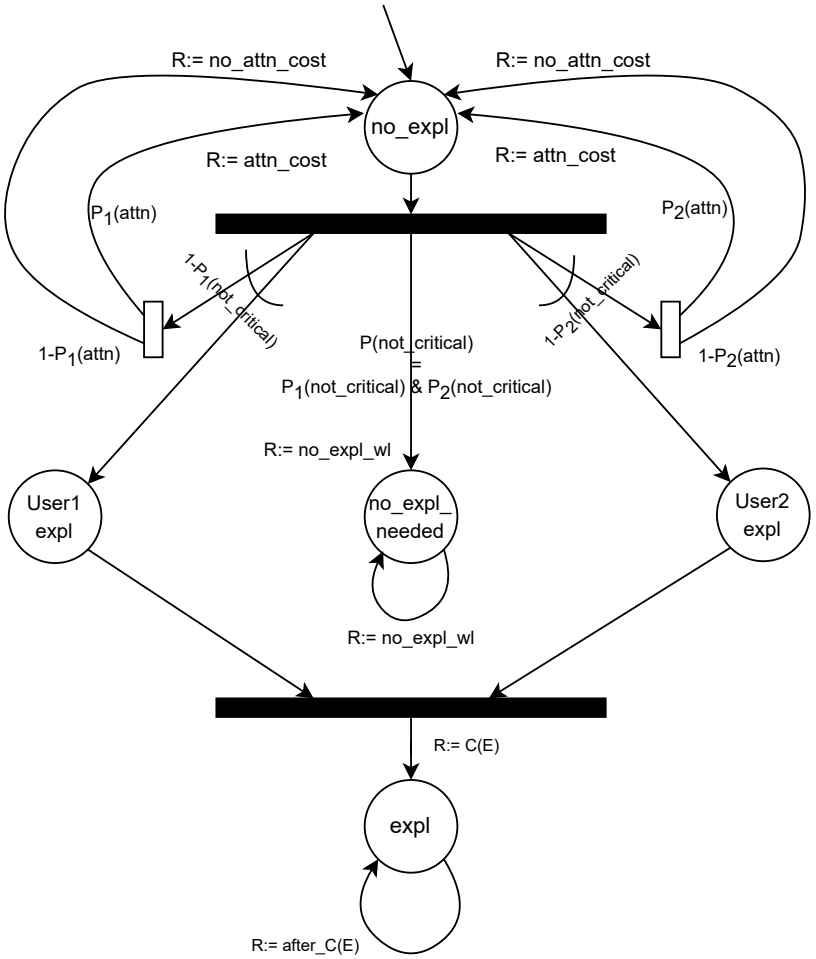


Figure 5.1: MDP representing the options for strategic player with parallel processes; to be combined be a product construction with the SEEV model in order to obtain the actual game graph

Figure 5.1 illustrates a detailed MDP that represents the various states and transitions for the strategic player, including parallel processes. At the beginning of the game, the strategic player starts in the *no_expl* state, indicating that no explanation is provided to the users. This initial state reflects a scenario where the player has not yet assessed the situation or determined the need for an explanation. From this point on, at every discrete time step (each step representing one second), the strategic player can transition to one

of the following states, with each transition governed by specific probabilities based on the users' assessment of the situation and the strategic player's decision-making process:

1. *no_expl_needed* State: The *no_expl_needed* state is reached when both users independently judge that the situation is not critical, as indicated by the joint probability $P(not\_critical)$. This means that the users perceive no immediate threat or need for intervention, and the strategic player, upon receiving this feedback, concludes that no explanation is necessary. Once the strategic player enters this state, they remain here until the end of the scenario. This state reflects a situation where both users are sufficiently confident in the current state of the environment, and there is no need to deviate from the initial plan or provide additional information.

2. *no_expl* State: The *no_expl* state represents a situation where the users perceive the scenario as critical (probabilities $1-P_i(not\_critical)$, where $i = 1, 2$ for the two users). However, no explanation is provided in this state.

   The reward structure in this state is nuanced and depends on whether users are actively paying attention or not. For instance, if users are engaged and attempting to deduce the vehicle's actions, their cognitive workload increases, potentially impacting their experience negatively. On the other hand, if users are not attentive, the strategic player may avoid penalization but risks leaving them uninformed. These reward dynamics make the decision to stay in this state a critical one. The details of these reward mechanisms are further elaborated in section 5.3.

3. *expl* State: The *expl* state is activated when the strategic player decides to provide an explanation, which can occur either proactively (based on the player's assessment of the scenario) or reactively (in response to user feedback indicating a lack of understanding). Regardless of whether the explanation is optimally timed, the decision to enter the *expl* state ensures that both users are informed. The timing of the explanation plays a significant role in the reward system, with optimal explanations leading to higher rewards. Notably, even if only one user expresses the need for clarification, the explanation is provided to both users. This ensures that all users are on the same page and prevents confusion from arising between them. Once the strategic player enters the *expl* state, they remain there for the remainder of the scenario, signalling that both users have received the necessary explanation to understand the vehicle's actions.

S	S'	Probability	R
no_expl	no_expl	P(not_critical) · P(attn)	attn_cost
no_expl	no_expl	P(not_critical) · (1-P(attn))	no_attn_cost
no_expl	expl	P(not_critical)	C(E)
no_expl	no_expl_needed	P(not_critical)	no_expl_wl
expl	expl	1	after_C(E)
no_expl_needed	no_expl_needed	1	no_expl_wl

Table 5.1: MDP Reward Structure

Reward in Figure	Reward/Cost Value
attn_cost	0.4
no_attn_cost	0.2
C(E)	0.3
after_C(E)	0.1
no_expl_wl	0.0

Table 5.2: Reward/Cost Values

5.3 Results and Discussion

The reactive game introduced in this paper serves two primary objectives: To determine the optimal timing for delivering an explanation to two users in a manner that minimizes the cognitive workload for both; and to identify the minimum expected cognitive workload across all potential explanation delivery strategies.

To achieve these goals, a reward structure was established for various state transitions, as depicted in table 5.1. These rewards and costs are grounded in probabilities and states derived from the model illustrated in fig. 5.1. The detailed values assigned to these rewards and costs are listed in table 5.2.

Since the model operates under a finite horizon framework, backward Bellman induction was utilized to compute the minimum cognitive workload induced by an attention strategy at any given time. The formula for the minimum workload (min_wl) is provided in eq. (5.5).

$$min\_wl_n^k = P_i(not\_critical) \cdot no\_expl\_wl + (1 - P_i(not\_critical)) \cdot expl\_wl_n^k \quad (5.5)$$

In this model, some instances may not require an explanation if the user has already assessed the surroundings or if the situation resolves on its own. The term $P_i(not\_critical),\ where\ i = 1, 2$, accounts for such instances by representing the prob-

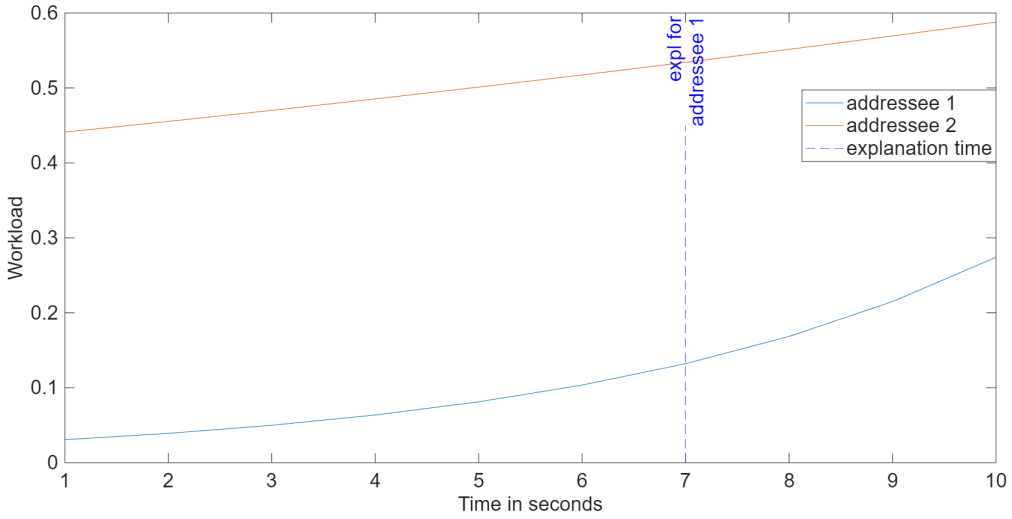
ability of a non-critical situation for each of the two users. If no explanation is necessary, a constant value no_expl_wl is used. The workload associated with providing an explanation, denoted as $expl\_wl_n$, depends on the timing of the event, represented by k, and the current time, n. The formula for $expl\_wl_n$ is provided in Equation 5.6.

$$expl\_wl_n^k = min \begin{cases} C(E) + (k-n) \cdot after\_C(E), \\ P_i(attn)_n \cdot (min\_wl_0^{k-n} + attn\_cost) \\ +(1 - P_i(attn)_n) \cdot (min\_wl_{n+1}^k + no\_attn\_cost) \end{cases} \tag{5.6}$$

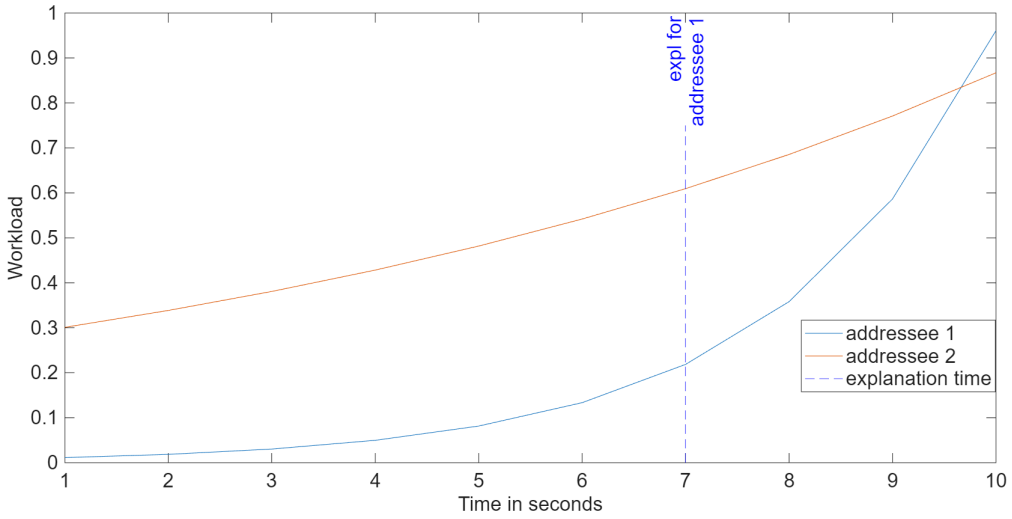
The $expl\_wl_n^k$ value represents the minimum between the cost of delivering an explanation and the cost of withholding one at time n, for a total scenario duration of k. This minimum value indicates the optimal strategy for timing the explanation, designed to minimize the expected cognitive workload. The cost of delivering an explanation involves both the immediate cost $C(E)$ and any subsequent costs after the explanation, represented as $after\_C(E)$. On the other hand, the cost of not delivering an explanation depends on the attention probability, which is determined through backward Bellman recursion. The attention probability for each user, $P_i(attn)\ \ where\ \ i = 1, 2$, is calculated using the SEEV model.

When attention is being paid, the variable $expl\_wl$ represents the workload cost associated with following an attention direction, denoted as $attn\_cost$, which is then combined with the backward recursion of the minimum workload. This recursion is adjusted for the remaining horizon, represented by $k - n$, to reflect the cumulative workload across the remaining time in the scenario. In contrast, when no attention is being paid, $expl\_wl$ represents the cost associated with the lack of attention, denoted as no_attn_cost, which is then combined with the backward recursion value of the minimum workload, capturing the potential cognitive workload resulting from not focusing on the task at hand.

Equation (5.5) and eq. (5.6) are interdependent, forming a set of mutually recursive relationships. By applying these equations, the optimal timing for delivering an explanation to each user is determined, based on the strategy that minimises the expected cognitive workload at any given point. Once the optimal times for each user are calculated, these times are compared, and the explanation is delivered to the user who requires it the soonest, thereby ensuring the most efficient reduction of cognitive workload for both users.



(a)

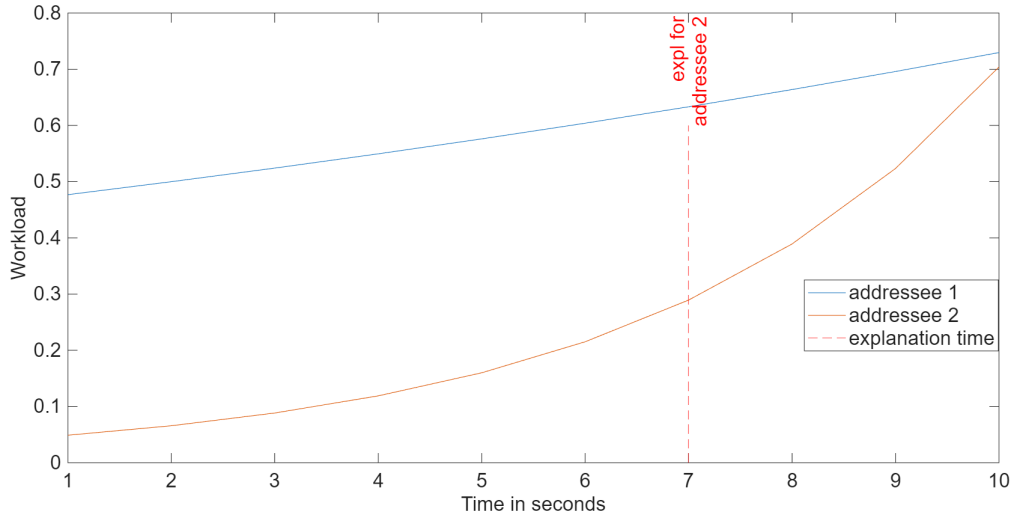


(b)

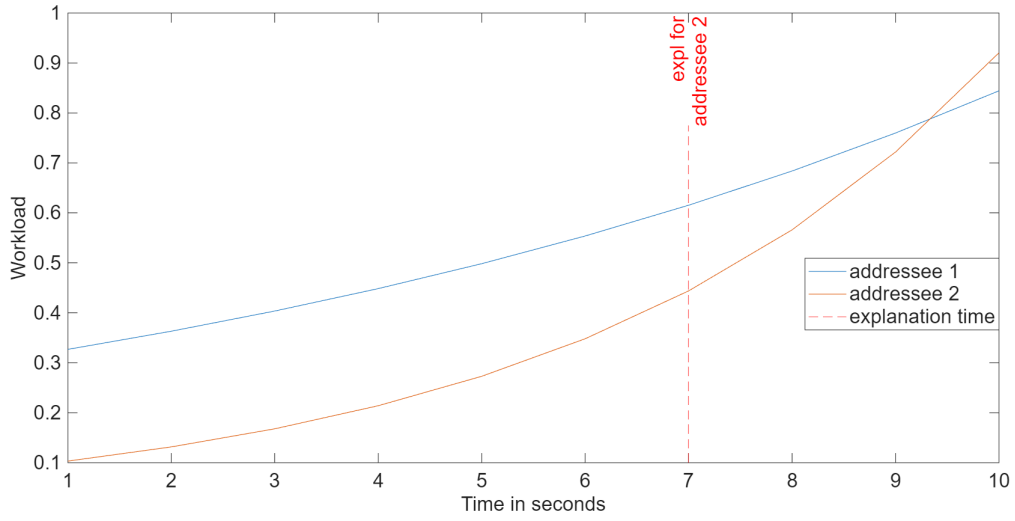
Figure 5.2: User/addressee 1 requires an explanation earlier

Figure 5.2, fig. 5.3, and fig. 5.4 present various graphs depicting the timing of explanations based on different scenarios. In these figures, the scenario occurs at 10 s, with varying expectations for users 1 and 2. Figure 5.2 shows two graphs where user 1 requires an explanation sooner than user 2. In the two graphs in fig. 5.3, we see the opposite, i.e. user 2 needs an explanation earlier than user 1. In fig. 5.4, both users require an explanation at the same time.

As observed from the graphs, the explanation is typically delivered 2 s or 3 s before the scenario concludes. Repeating the experiment with different scenario end times reveals



(a)



(b)

Figure 5.3: User/addressee 2 requires an explanation earlier

that when the scenario duration is 2 s or less, providing an early explanation does not significantly reduce the cognitive workload. This finding is consistent with prior research on explanation timing for single users that we showed in chapter 3.

Since the model is implemented using MATLAB and relies on Backward Induction, even a small increase in scenario duration (such as one second) leads to a noticeable increase in computational demands. For scenarios lasting longer than 25 seconds, the computation time for determining the optimal explanation timing exceeds 2.5 seconds, making MATLAB an inefficient platform for implementing this model in such cases.

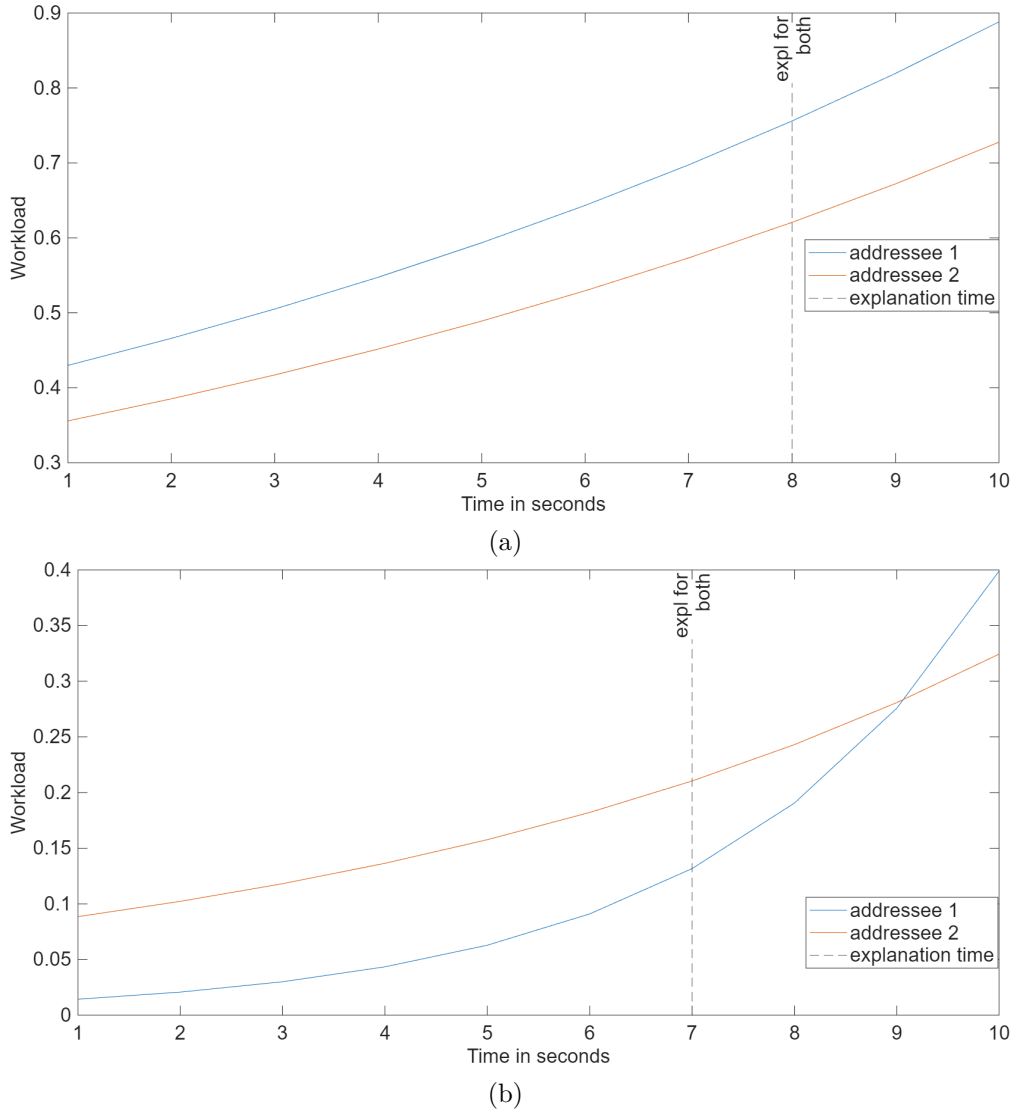


Figure 5.4: Both users require an explanation at the same time

5.4 Chapter Summary

In this chapter, we aimed to determine the optimal timing for providing an explanation in autonomous systems involving multiple users, using a game-theoretic approach. Our findings demonstrate that explanation timing strategies can be effectively adapted to accommodate the diverse needs of different users. While our models provide valuable insights, they are constrained by the assumptions made for the SEEV model and the reward structure and values shown in Tables 5.1 and 5.2. These rewards/costs are

currently based on educated estimates intended to demonstrate the technology, but they lack empirical psychological validation.

In the next chapter, we will see how to expand this model to multi-step explanations for multiple users.

6 Multi-Step Explanations for Multiple Users

Contents

In chapter 3 and chapter 4, we dealt with providing either an atomic or multi-step explanation to a single user, and in chapter 5, the optimal time for providing an atomic explanation for multiple users was addressed. In this chapter, we now focus on combining the knowledge of the previous chapters and determining the optimal time for providing multi-step explanations for multiple subjects. The contents of this chapter are based on our submitted work [BFS25].

Our contributions in this chapter are as follows:

1. we propose a model-based approach to optimizing explanation timings based on the mental workload and attention of multiple users
2. we implement a game-theoretic framework to simulate the attention and workload of the users and to optimise the timing of AV
3. we demonstrate the feasibility of multi-step explanation timing for multiple users in AV settings

This chapter is divided as follows: in section 6.1, we discuss the existing works in this field. We describe an example scenario to better understand the need for our work in section 6.2.The implementation of SEEV in a reactive game is explained in section 6.3. The results of the reactive game are discussed in section 6.4, and we finally conclude this chapter by summarizing the key contributions in section 6.5.

6.1 Existing Research

The intersection of multi-step explanations and their delivery to multiple users remains an underexplored area of research. Stansberry's dissertation [Sta12] examines the flow of information in online communities, focusing on the critical role of influencers in multi-step dissemination processes involving diverse audiences. Similarly, Finzel et al. propose a user-centric framework for constructing multi-level and multi-modal explanations [Fin+21]. Their process-oriented approach enables users to interactively explore explanations at various levels of detail, fostering greater comprehension and engagement. Beyond these two studies, we found no other research that explicitly addresses both multi-step explanations and their application to multiple users.

6.2 Example Scenario

Explanations typically involve several key pieces of information, and the timing of when these details are shared can fluctuate based on the amount of information available at any given moment. In situations where the information is either incomplete or still unfolding, it becomes especially important to carefully consider when to convey the relevant details. This decision is critical because it significantly influences the clarity, comprehensiveness, and overall effectiveness of the explanation, ultimately determining how well it achieves its intended goal.

AVs, for instance, often encounter situations where they need to assess and respond to dynamic traffic conditions, such as merging onto a highway (see fig. 6.1).

Example 6.2.1. Consider an AV entering a highway via an on-ramp. Initially, its sensors detect fast-moving vehicles in the adjacent lane, but there is uncertainty regarding their exact speeds and intentions. In this scenario, the AV must decide whether to provide the partial information it currently has or wait until more data becomes available.

In such cases, if the AV immediately informs the passenger that it is preparing to merge, but lacks complete information, this premature explanation might lead to increased cognitive load for the passenger, making them anxious or unsure about the vehicle's next move. On the other hand, if the AV waits until it has gathered more comprehensive data, such as a precise gap analysis and speed matching, it can offer a more confident, complete explanation of its decision. This approach not only reduces unnecessary cognitive strain for the passenger but also helps maintain their situational awareness. However, delaying the explanation for too long could backfire, causing confusion or anxiety as the passenger remains uncertain about the vehicle's intentions [Gu+20].

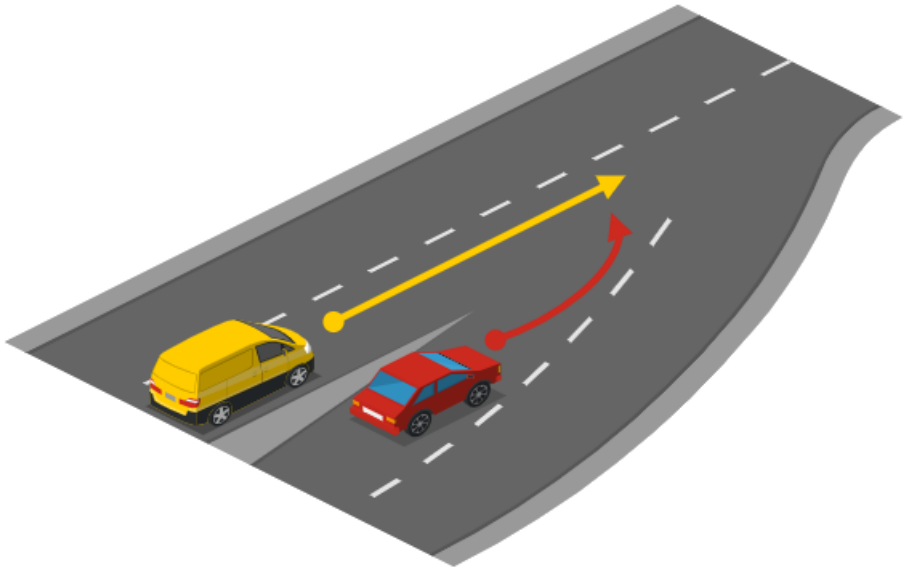


Figure 6.1: AV lane merging example: AV needs to merge onto a lane with oncoming traffic

This dilemma highlights a key challenge faced by AVs and other interactive systems: the need to manage the timing and content of multi-step explanations. The decision to provide partial versus complete explanations is not merely about the availability of information but also involves understanding the mental workload and attention demands placed on users [Kan00; Kul+13b]. In multi-user settings, where attention is often divided, determining the optimal timing of explanations becomes even more complicated, as the system must account for the cognitive load and attention of all individuals involved.

6.3 Multi-Player Game

This chapter aims to optimise the timings of the provision of multiple explanations for multiple users. And we do that by using the attention model SEEV to determine the mental workload of the users. For ease of implementation, we only consider two users in the AV.

The modified SEEV formula, as presented in eq. (2.3), in conjunction with the SEEV two-users model outlined in section 5.2.1, is employed to assess and determine the attentional focus of the users in this chapter. This approach enables a comprehensive analysis of how attention is distributed across multiple users within the given context.

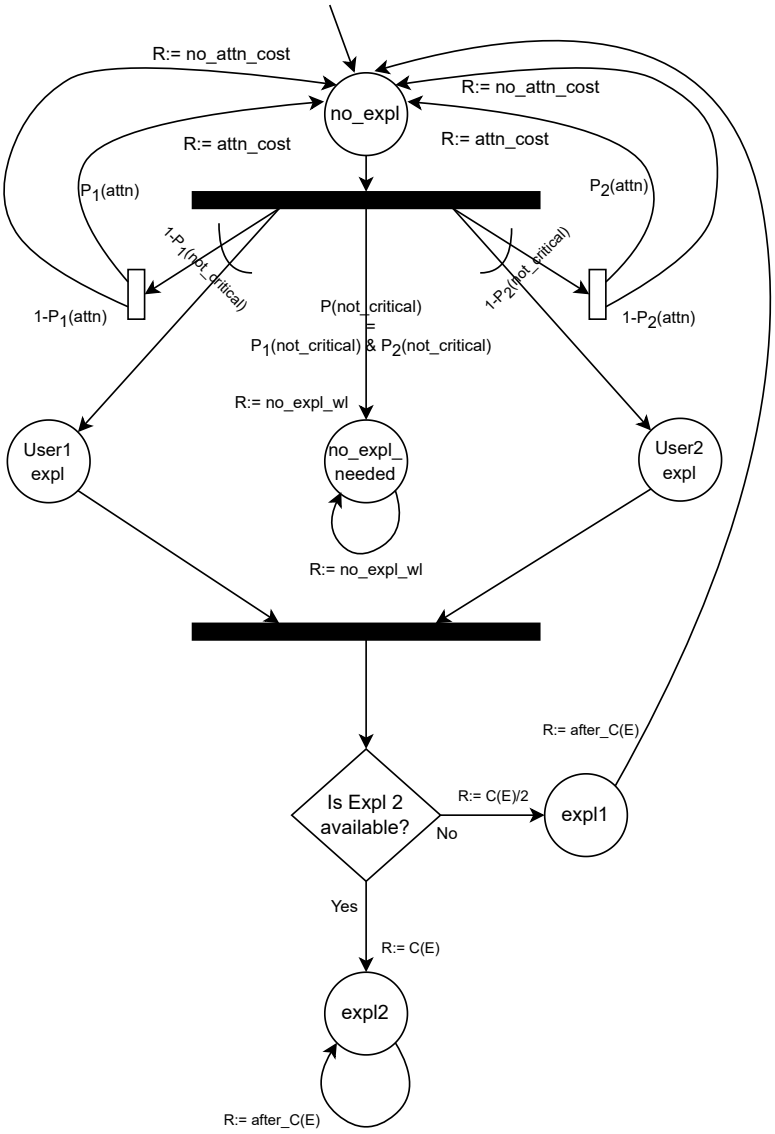


Figure 6.2: MDP representing the options for strategic player with parallel processes and multi-step explanation; to be combined be a product construction with the SEEV model in order to obtain the actual game graph

Since explanations involve multiple pieces of information, their timing delivery depends on the availability of relevant details. Consider the highway merging scenario from example 6.2.1, where an AV must merge into a lane with ongoing traffic. To ensure safety amid uncertainty, the AV begins decelerating, but it only has partial information—such as detecting a fast-approaching vehicle.

This raises a key question: Should the AV share incomplete information immediately or wait for a full understanding? Early explanations provide context but may increase cognitive load if users try to fill in gaps. Delayed explanations, meanwhile, risk uncertainty as users form their own assumptions.

The modified SEEV model, shown in eq. (2.3), is applied as a dynamic factor impacting workload within a Markov Decision Process (MDP) game. In this game, decisions are based on optimal timing for presenting explanations. Using MATLAB [MAT24], we calculated the reactive strategy for delivering explanations. The SEEV-based game starts k seconds before the scenario begins and ends when the scenario concludes.

Figure 6.2 presents a MDP illustrating various states and transitions for the strategic player, which operates in parallel processes. Initially, the player starts in the *no_expl* state, indicating no explanation is provided.

At each one-second time step, the strategic player may transition to one of the following states based on certain probabilities:

1. *no_expl_needed:* Entered only if both users consider the situation non-critical (i.e., $P(not\_critical)$). Once here, the player remains in this state until the scenario ends.

2. *no_expl:* Entered when both users view the situation as critical (1 - $P_i(not\_critical)$ where $i = 1, 2$). Rewards vary based on whether the user is attentive, as detailed in section 6.4.

3. *Is expl*2 *available?:* Reached when the strategic player decides to deliver an explanation. It is first checked if a complete explanation ($expl2$) already exists. If so, and if either user needs an explanation, both receive it, and the player stays in this state until the scenario ends. If the complete explanation is not yet available, then only a partial explanation ($expl1$) is provided and the strategic player returns to the *no_expl* state.

6.4 Model Output

The objectives of the multi-user reactive SEEV game presented in this chapter were expanded to address three main goals:

1. *Evaluating Explanation Structure and Timing:* Identify whether explanations should be delivered in two parts, given that the Autonomous Vehicle (AV) determines

information asynchronously in two segments. If a segmented explanation is necessary, the study aims to identify the optimal timing for each part, minimizing the cognitive workload of the human users.

2. *Optimizing Explanation Timing for Cognitive Workload Reduction:* Establish the optimal timing for delivering explanations to both users to reduce cognitive workload.

3. *Minimizing Expected Cognitive Workload Across Presentation Strategies:* Determine the minimum cognitive workload for each user across various presentation strategies (either only a complete explanation or both partial as well as complete explanations).

To achieve these objectives, several methodological steps were followed:

1. *Assigning Rewards to State Transitions:* The study utilizes a reward-based structure to assess different state transitions within the game, as shown in fig. 6.2. Transition states were assigned specific rewards or costs, outlined in table 6.1, which served as the basis for assessing optimal explanation strategies.

2. *Defining Probability-Based Reward/Cost Values:* The reward/cost values applied to various probabilities are provided in table 6.2. These values are essential for evaluating transitions where a particular cognitive workload is incurred, depending on whether a critical explanation is necessary or can be omitted.

3. *Application of Backward Bellman Induction:* Given the finite horizon model of this game, backward Bellman induction (BI) was used to calculate the minimal cognitive workload. This method allows iterative calculation, ensuring that the minimum workload is achieved through the attention strategy at any given point in time. The formula for minimum workload, denoted min_wl, is expressed as:

$$min\_wl_n^k = P_i(not\_critical) \cdot no\_expl\_wl + (1 - P_i(not\_critical)) \cdot expl\_wl_n^k \tag{6.1}$$

Here, $P_i(not\_critical)$ represents the probability that no critical situation arises, eliminating the need for an explanation. When an explanation is unnecessary, a constant cognitive workload, no_expl_wl, is assigned. Otherwise, the workload for providing an explanation, $expl\_wl$, depends on the time of the scenario (denoted k) and the current point in time (denoted n).

S	S'	Probability	R
no_expl	no_expl	(1 - $P_1(not\_critical)$) ·$P_1(attn)$	attn_cost
no_expl	no_expl	(1 - $P_1(not\_critical)$) ·(1 - $P_1(attn)$)	no_attn_cost
no_expl	no_expl	(1 - $P_2(not\_critical)$) ·$P_2(attn)$	attn_cost
no_expl	no_expl	(1 - $P_2(not\_critical)$) ·(1 - $P_2(attn)$)	no_attn_cost
no_expl	User1 expl → is expl2 available? → expl1	(1 - $P_1(not\_critical)$)	$C(E)/2$
no_expl	User2 expl → is expl2 available? → expl1	(1 - $P_2(not\_critical)$)	$C(E)/2$
no_expl	User1 expl → is expl2 available? → expl2	(1 - $P_1(not\_critical)$)	C(E)
no_expl	User2 expl → is expl2 available? → expl2	(1 - $P_2(not\_critical)$)	C(E)
no_expl	no_expl_needed	P(not_critical)	no_expl_wl
expl1	no_expl	1	after_C(E)
expl2	expl2	1	after_C(E)
no_expl_needed	no_expl_needed	1	no_expl_wl

Table 6.1: MDP Reward Structure for multi-users, multi-step explanations

4. *Considering the Cost of Explanation versus Non-Explanation:* The workload associated with providing an explanation, $expl\_wl_n^k$, shown in eq. (6.2), represents the minimum cost between offering an explanation and refraining from doing so, given the scenario's duration. This trade-off is calculated based on:

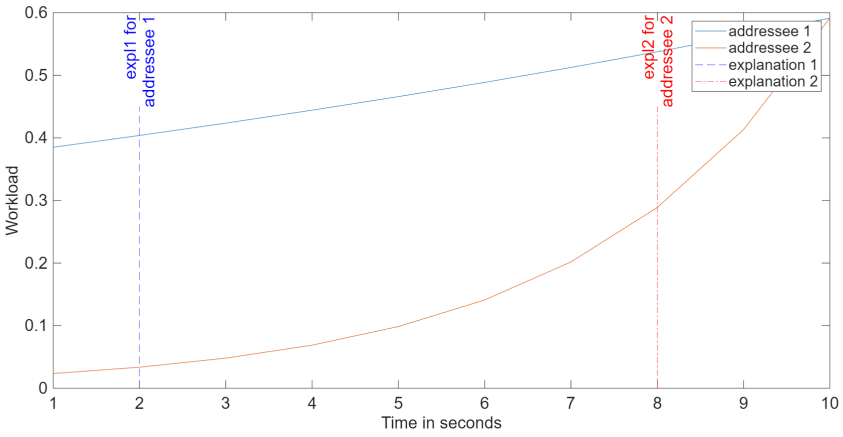
$$expl\_wl_n^k = min \begin{cases} expl\_cost, \\ P_i(attn)_n \cdot (min\_wl_0^{k-n} + attn\_cost) \\ +(1 - P_i(attn)_n) \cdot (min\_wl_{n+1}^k + no\_attn\_cost) \end{cases} \tag{6.2}$$

$$expl\_cost = \begin{cases} C(E) + (k-n) \cdot after\_C(E) & expl2 \ is \ available \\ 0.5 \cdot C(E) + (k-n) \cdot after\_C(E) & otherwise \end{cases} \tag{6.3}$$

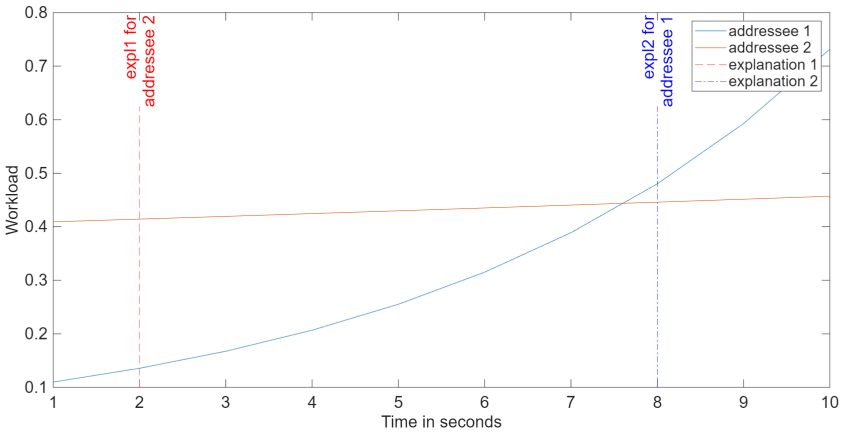
- *Explanation Cost:* Denoted by $expl\_cost$, it includes the direct cost of the explanation, $C(E)$, and any additional costs post-explanation ($after\_C(E)$). This is given by the eq. (6.3) and if $expl2$ is already available, then the cost of

Reward in Figure	Reward/Cost Value
attn_cost	0.4
no_attn_cost	0.2
C(E)	0.3
$C(E)/2$	0.15
after_C(E)	0.1
no_expl_wl	0.0

Table 6.2: Reward/Cost Values



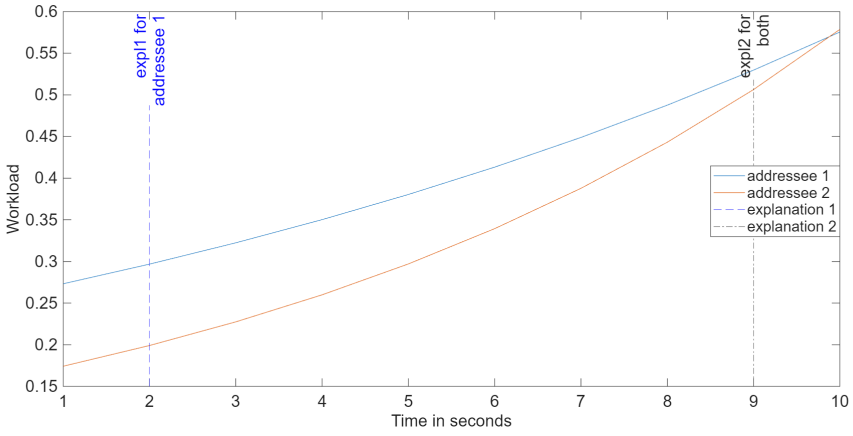
(a)



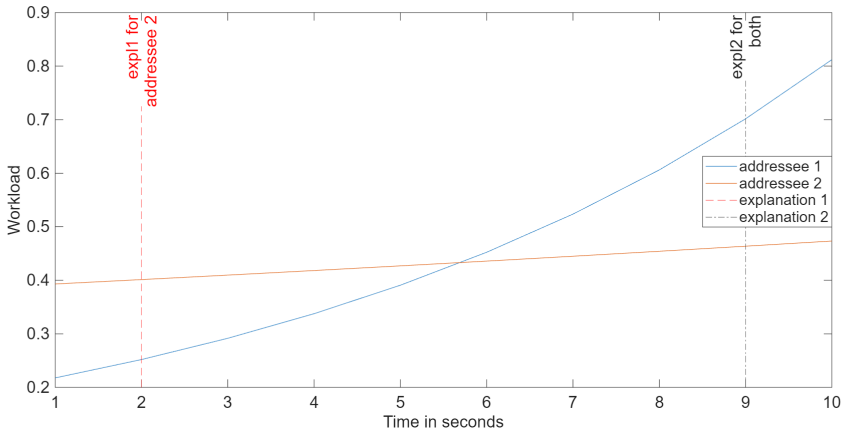
(b)

Figure 6.3: One of the users requires a partial explanation and the other user requires a complete explanation

providing an explanation is $C(E)$. If only a partial explanation is provided, then half of $C(E)$ is the cost of the explanation.



(a)

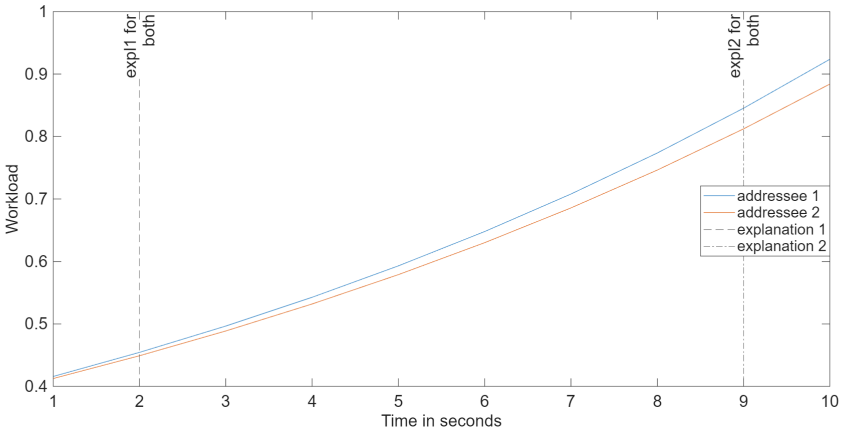


(b)

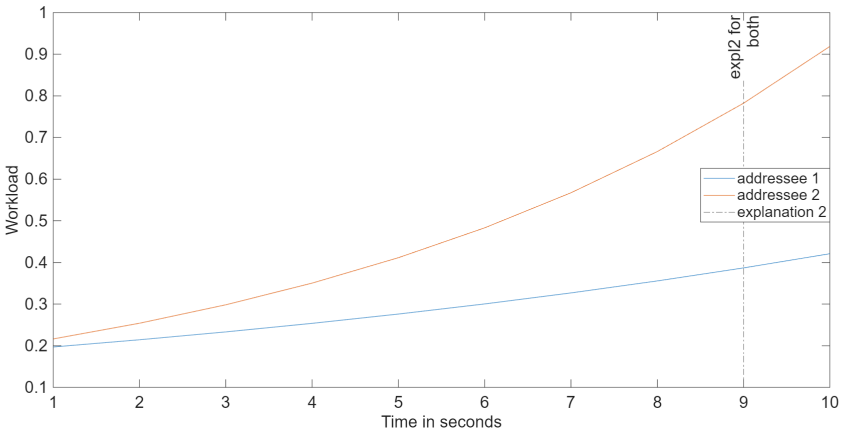
Figure 6.4: One of the users requires a partial explanation and both the users require a complete explanation

- Non-Explanation Cost: Varies based on the probability of user attention, derived using backward Bellman recursion. This probability, $P_i(attn)$, is influenced by the SEEV model.

5. *Recursive Workload Calculation:* Equation (6.1) and eq. (6.2) are mutually recursive, providing a systematic approach to calculating the optimal timing for explanations, both as single and potentially multi-part explanations. By comparing these values, it becomes possible to determine whether and when each part of an explanation should be delivered to each user to minimize their cognitive workload.



(a)

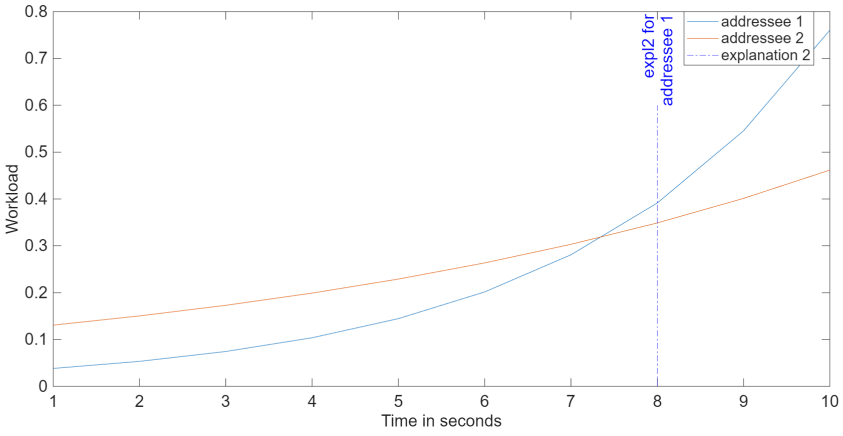


(b)

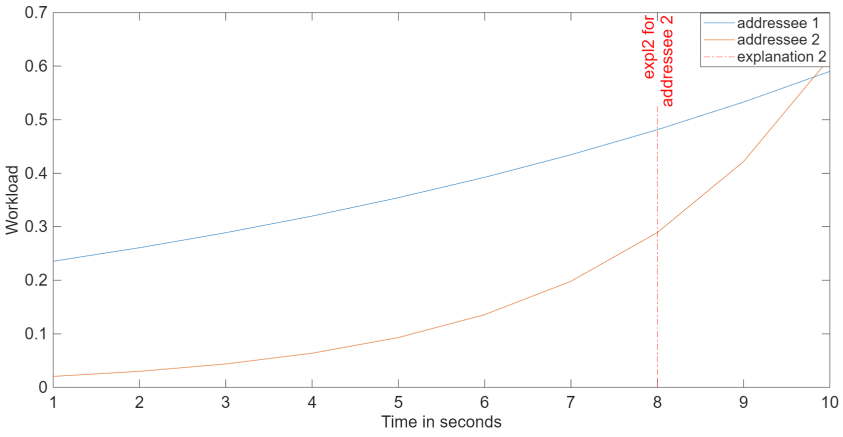
Figure 6.5: Both users require either a complete explanation or both a partial and a complete explanation

Based on all the methodological steps mentioned above, we conducted our SEEV reactive game in MATLAB. Further, we'll discuss some of the key results we observed when running the game in a scenario that lasted for 10 s.

Figure 6.3 shows two graphs, which required both partial and complete explanations. The partial explanation was given due to the mental workload of one user, while the complete explanation was provided based on the workload of the other user. In the graphs in fig. 6.4, the initial cognitive workload is higher for one of the users, and thus the explanation timing was based on this user's workload. However, when a complete explanation is available, the cognitive workload of both users is high thus both users



(a)



(b)

Figure 6.6: Only one of the users requires a complete explanation

need the complete explanation. In fig. 6.5, both users require either both explanations or only the complete one. Lastly, in some cases, as shown in fig. 6.6, neither of the users needs a partial explanation and only one of the users needs a complete explanation.

An interesting observation from these figures is that when a partial explanation is needed, it is always provided at 2 s as this timing minimizes the cognitive workload for users the most. The complete explanation is provided at 8 s if only one user requires it. In all cases where both users need a complete explanation, it is provided at 9 s.

6.5 Chapter Summary

In this chapter, we focused on integrating frameworks for multi-step explanations for a single user and atomic explanations for multiple users. We expanded the previous SEEV reactive game to determine the optimal times for delivering partial and complete explanations to two users. The results revealed that even when the partial explanation is provided based on the cognitive workload of one user (U1), the other user (U2) may require the complete explanation before U1 does.

7 Case Study: Interactive Explanation Timing Game

Contents

Up to this point, we have explored the implementation of models designed to determine the optimal timing for providing explanations. These models were built on a set of predefined assumptions (the reward/cost values), which served to streamline the development process by reducing the complexity of the underlying logic. While these assumptions helped facilitate the model's creation, it is essential to rigorously evaluate the model's performance in a more practical and user-centred context. To achieve this, we designed and conducted a user study in the form of an interactive game, allowing us to assess the model's effectiveness in real-time, user-driven scenarios. The content of this chapter is based on our previously published work in [Bai+25]

This chapter is divided as follows: in section 7.1, we give an overview of existing studies/work on the determination of optimal timing for providing explanations. Then we discuss the various steps involved in the study design in section 7.2. Section 7.3 talks about the ethical considerations that were done for this study. We describe the process of our data collection in section 7.4. In section 7.5, we analyse the results of our study

in detail. In section 7.6, we discuss the implications and limitations and our work. And finally, in section 7.7, we give a summary of this chapter.

7.1 Related Work

Our work sits at the intersection of explanation timing and mental workload. In this section, we provide an overview of both areas, discussing key concepts, relevant theories, and prior research.

Explanation timing in the domain of autonomous driving. Recent research highlights the importance of providing explanations in Autonomous Vehicles (AVs) to enhance the user's trust and situational awareness. Kim et al. [Kim+24] introduced TimelyTale, a multimodal dataset that identifies passengers' needs for explanations in automated vehicles and predicts when context-aware explanations are most beneficial. Additionally, another study by Kim et al. demonstrated that visualizing an AV's perception improves passenger experience without increasing cognitive load. They also found that timing of explanations based on traffic risks effectively mitigates information overload [Kim+23].

Explanation timing: Before or after an event? Cognitive science research on reaction —time tasks—tasks that are repetitive and temporally predictable— has shown that stimuli presented earlier than expected improve accuracy, whereas those arriving later than expected increase the likelihood of errors [GRE01].

Beyond reaction-time tasks, Chen et al. [CLS24] examined how explanation timing (pre-action, post-action, both, or none) affects user trust, comprehension, and satisfaction with AI systems. They found that pre-action explanations help users anticipate biases in AI, while post-action explanations enhance retrospective understanding. Combining both led to better trust calibration, aligning user expectations with the AI's capabilities and limitations.

The impact of explanation timing on other research areas. While explanation timing intersects with areas like User eXperience (UX) Engineering [AD13], research on event timing within UX engineering falls outside the scope of this paper. However, recent studies [Det+24] indicate that well-designed explanations can enhance user experience, whereas poorly implemented explanations may introduce risks. This highlights the necessity of studying explanation timing as a foundational aspect of integrating explanations into UX design.

For example, providing timely explanations for an autonomous system's actions can enhance user trust and situational awareness. Elbitar et al. [Elb+21] explored how

the timing and rationale of runtime permission requests influence user decisions and their evaluation of those decisions. Their findings offer valuable insights for enhancing permission request strategies and overall user experience.

While these studies highlight the critical role of explanation timing, none have specifically investigated how to align explanation timing with a user's cognitive workload. To address this gap, our research focuses on identifying the optimal timing for explanations by conducting a user study that evaluates the cognitive effort required for users to process and comprehend explanations.

7.2 Study Design

The study was designed as a two-part game to evaluate the role of timing in explanation delivery. The game was developed using GDevelop 5 [Riv+21] —an open-source, no-code game development platform that allows users to create 2D, 3D, and multi-player games without needing extensive programming knowledge. This choice of engine was motivated by its flexibility and ease of use, making it suitable for prototyping and rapid iteration. Additionally, GDevelop supports a wide range of functionalities that are necessary for this study, such as customizable logic events, user interaction tracking, and data collection features.

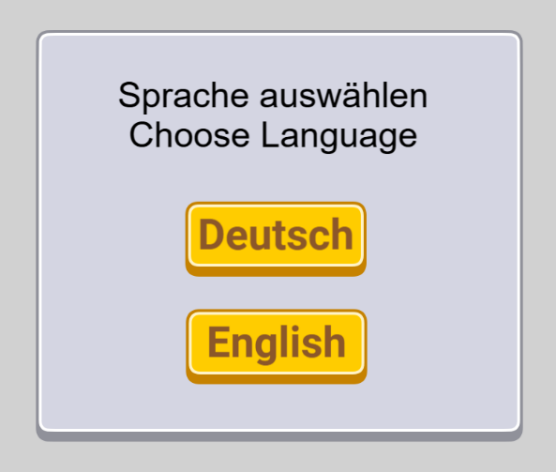


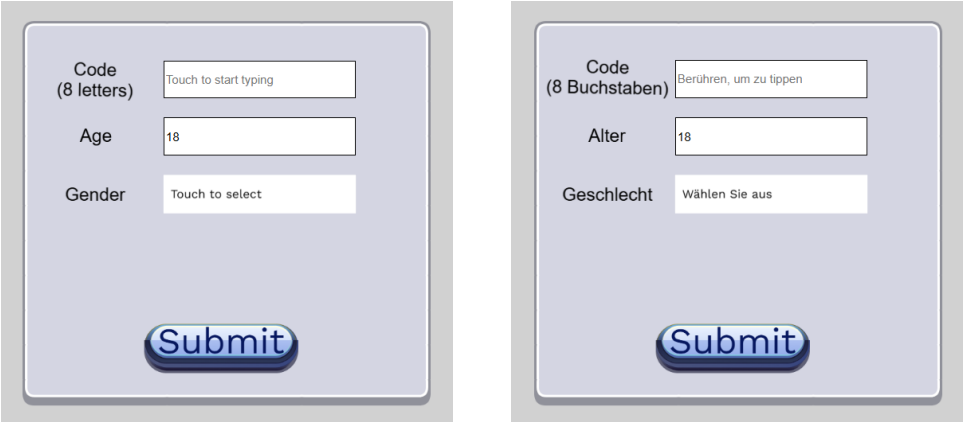
Figure 7.1: Initial language selection page for the participants

Although the game is currently designed for offline use, the architecture was built with future scalability in mind, meaning it could be easily adapted for online deployment. An online version of the game could enable a broader range of participants to engage in the study, providing richer and more diverse data. Moreover, an online platform could

facilitate real-time data collection and remote updates to the game, allowing for more dynamic adjustments based on user behaviour.

At the start of the study, participants were asked to select their preferred language. As shown in fig. 7.1, they could choose between Deutsch (German) and English. Once selected, all subsequent instructions and game details were presented in the chosen language, ensuring clarity and accessibility.

Following language selection, participants provided demographic information, which was linked to a unique codeword assigned to each individual. Ethical considerations related to the data collection in this step are discussed later in the chapter. Figure 7.2 displays the demographic data entry page in both English and German.



(a) English Version (b) German Version

Figure 7.2: Demographic data entry page where participants provide age, gender, and their assigned codeword

Our study's goal is to determine the time taken by users to comprehend a simple explanation within a gamification setting. In our meaning, a simple explanation refers to explanations that exert minimal workload on the human. Explanations containing one to two words would be considered simple explanations. Examples for these could be "Stop!", "Turn Left" etc..

To enable the participants' understanding of these simple explanations, they were given a detailed explanation of how the game works beforehand. During the game, they received brief, concise instructions, which were simplified versions of the initial, more detailed explanation. This two-step approach is based on Krull's framework [Kru99], which argues that for a layperson to understand a concept effectively, it should first be thoroughly explained, followed by specific instructions outlining the next steps.

By embedding the study in a game format, we created a controlled, yet immersive and engaging, environment. This approach allowed us to simulate real-world scenarios where timely explanations are critical, enabling us to observe their impact on user performance, adaptability, and mental effort in comprehending and responding to explanations in real time. This was accomplished by measuring participants' reaction times and decision accuracy as they responded to instructions under varying cognitive demands.

The study was structured into three parts:

1. *Reaction Time Determination:* In this introductory task, participants are presented with a sequence of instructions corresponding to different colour names, each mapped to specific arrow keys (e.g., pressing the left arrow key for "red" or the up arrow key for "blue"). A colour is displayed on the screen, prompting the participant to press the associated arrow key as quickly as possible. This task establishes a baseline for participants' reaction times and their ability to associate visual stimuli with motor actions.

2. *Reactive Game:* The second task builds upon the baseline task, introducing more dynamic and cognitively demanding gameplay. Participants navigate lanes to collect coins of a specified target colour, adjusting their actions based on dynamically changing instructions. The game begins with a practice round, allowing participants to familiarize themselves with the mechanics and the goal of the task. Following this, participants proceed to the experimental round, during which performance data is collected. Throughout the game, instructions indicating the target coin colour are periodically updated, requiring participants to react promptly while avoiding distractions from non-target coins or irrelevant stimuli. Players need to use the left and right arrow keys to navigate lanes and collect the correct coins. This task assesses reaction times, decision accuracy, and participants' ability to process changing instructions under conditions of increased cognitive demand.

3. *Subjective Evaluation:* After completing the gameplay tasks, the participants were asked to fill out the NASA Task Load Index (NASA-TLX) form. NASA-TLX is a widely used tool for assessing perceived workload, where participants rate tasks across six dimensions: mental demand, physical demand, temporal demand, effort, performance, and frustration. Participants rate their experiences on each dimension, providing quantitative insights into the difficulty and cognitive burden of the tasks. These scores are used to gauge user experience and identify specific aspects of the tasks that participants found challenging or stressful, complementing the objective performance data collected during the gameplay.

At the end of each game, we measured how quickly players reacted to the explanations given on-screen and how effectively they followed the changing instructions. This allows us to assess the impact of explanation timing on decision-making within a dynamic environment.

We provide details for these three parts of the study in the next Sects. 7.2.1-7.2.3.

7.2.1 Reaction Time Determination

In the first phase of the study, participants underwent a task designed to measure their reaction times to different visual instructions. This phase consisted of two rounds: a test round and the actual experiment round.

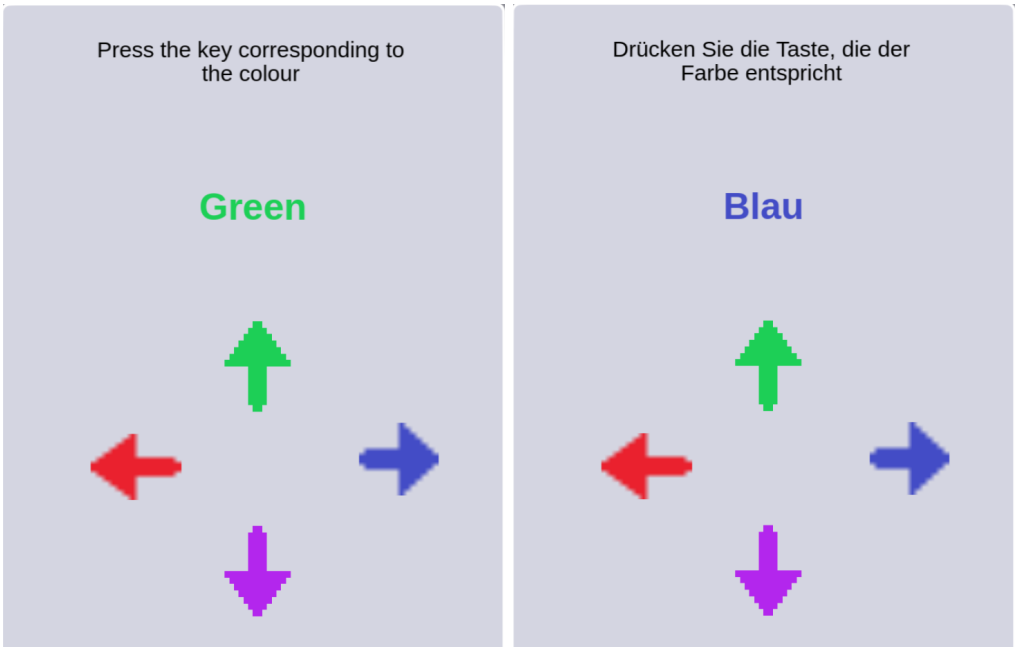
Test Round

In the test round, participants were introduced to the game mechanics and instructions. This round served as a warm-up, allowing users to familiarise themselves with the task and to understand what to expect in the subsequent rounds. No data from the test round were used in the final analysis, as it was intended solely to help participants get comfortable with the task.

This round consisted of instructions (colour name) being displayed on the screen, and these colour names were mapped to specific arrow keys (up, down, left, and right). The arrow keys with the colour mapping were shown directly below the instructions. The colour names (instructions) were displayed for 5 s before it changed, i.e. the participant had 5 s time to comprehend the instruction and press the correct arrow key direction. Fig. 7.3a shows an example of this task in the English version, while Fig. 7.3b presents the corresponding German version. The distinct mapping of each colour to a specific direction was intended to test participants' ability to process visual stimuli and translate them into accurate motor responses, providing a foundational measure of their reaction and comprehension capabilities.

Experimental Round

The second round served as the experimental phase, during which data were systematically collected for analysis. This phase was divided into four distinct sub-rounds, each corresponding to one of the four arrow key directions: up, down, left, and right. Similar to the test round, the instruction (colour name) was displayed for 5 s before the instruction was removed. There was 2 s pause between the sub-rounds.



(a) English Version (b) German Version

Figure 7.3: A snapshot of an instance in the reaction time determination game

In each sub-round, the colour associated with a specific arrow key was randomised, ensuring that participants could not rely on memorised colour-direction mappings from previous rounds. This randomisation was an essential feature, as it prevented learned behaviours and kept participants actively engaged with the task.

During each trial, when a specific colour appeared on the screen, participants were required to press the arrow key corresponding to the colour displayed on the screen as quickly as possible. Upon pressing the correct key, a confirmation message—"key pressed"—was displayed to provide immediate feedback. To increase the cognitive complexity of the task, an additional challenge was introduced: for e.g., when a colour associated with the *up* direction appeared, a shaking arrow pointing in a different direction (e.g. right) was simultaneously displayed. This visual distraction was meant to add an extra layer of difficulty, requiring participants to focus more intensely and resist the misleading cue.

The inclusion of the shaking arrow was designed to simulate heightened cognitive demand, allowing the study to evaluate reaction times under conditions of increased mental workload. Reaction times and error rates were recorded across all sub-rounds, providing valuable insights into how explanation timing and distractions influenced participants' performance and decision-making processes.

7.2.2 Reactive Game

The second phase of the study involved a more complex, reactive game designed to test participants' ability to follow instructions while managing multiple visual elements. Like the first phase, this phase consisted of two rounds: a test round to help participants understand the game mechanics and an experimental round in which data were collected.

Test Round

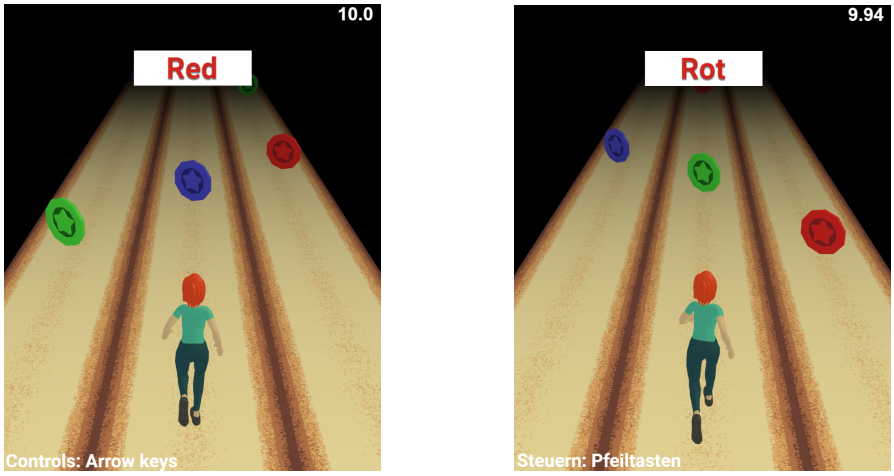
In the test round, participants were introduced to the mechanics of the reactive game through a practice session. The game involved using an avatar to collect coins of specific colours displayed on three separate lanes. This practice round ensured that participants were comfortable with the controls and could accurately interpret the instructions before advancing to the experimental phase. Importantly, no data were recorded during this round, as its primary purpose was to prepare participants for the main experiment.

The test round lasted for 45 s, with each instruction (a colour name) displayed for 2 s. Participants were given a total of 10 s to collect the coin of the correct colour after the corresponding instruction was shown, providing ample time to navigate and understand the task mechanics. This setup allowed participants to develop familiarity with the task flow, reducing potential confusion or errors during the experimental phase.

Experimental Round

In the experimental round, participants collected coins matching the target colour displayed at the top of the screen. As they progressed, the target colour changed periodically, requiring them to adjust their strategy and actions dynamically. The game environment during this phase can be seen in Fig. 7.4a, which shows the English version, and in Fig. 7.4b, which shows the German version. They provide a clear view of the game layout and instructions.

In this phase, the participants controlled an avatar using the left and right arrow keys to move between the three lanes. Coloured coins appeared randomly in each lane, and participants were tasked with collecting coins that matched the target colour displayed prominently at the top of the screen (see Fig. 7.4). As illustrated in the figures, the target colour ("Red" or "Rot") was clearly indicated in the centre, serving as a guide to help participants focus on the correct coin. The primary objective was to collect as many coins of the designated colour as possible within the allotted time. This helped us test their ability to follow instructions and their decision-making under time pressure.



(a) English Version (b) German Version

Figure 7.4: A snapshot of the reactive game for different language settings

The game was played over a total duration of 85 s. Each instruction, indicating the target colour, was displayed for 2 s. After each instruction, a 9 s interval followed, during which participants collected coins matching the previously displayed target colour. This interval allowed participants to focus on the task without continuous interruptions, while also testing their ability to remember and apply the most recent instruction.

To increase complexity, the game frequently updated the target colour. This required the participants to dynamically adapt their actions in response to new instructions. As the game progressed, coins in various colours (e.g., red, blue, green) appeared in random lanes, compelling participants to make quick decisions about which lane to move into based on the updated target colour.

Each time a participant successfully collected a coin of the correct colour, they continued navigating the game until the next target colour instruction appeared. The scoring system rewarded participants based on the accuracy of their coin collections, with points granted for each correct coin gathered. Reaction times, accuracy rates, and total scores were recorded to provide comprehensive performance data for each participant.

This phase of the study offered valuable insights into how participants responded to dynamically changing instructions, remembered prior instructions, and managed their cognitive workload in a fast-paced, decision-intensive environment. The data collected helped evaluate participants' ability to process new information quickly, adjust their actions accordingly, and perform effectively under varying task conditions.

7.2.3 Subjective Evaluation Using NASA Task Load Index

The final part of the study was user feedback. The participants were asked to fill out the NASA Task Load Index (NASA-TLX) form [HS88][1]. NASA-TLX, developed by Hart and Staveland, is a widely utilized subjective workload assessment tool that provides an empirical method for quantifying subjective perceptions of workload in various task environments. It was designed to evaluate perceived workload across six dimensions.

The six dimensions in the NASA-TLX capture different facets of workload that impact task performance and cognitive load.

- *Mental Demand:* Evaluates the level of cognitive effort required to complete the task. It reflects how mentally challenging participants found the activity and whether it required sustained focus.
- *Physical Demand:* Assesses the physical effort exerted during the task. This dimension gauges whether participants felt any strain or fatigue from the physical aspects of the activity.
- *Temporal Demand:* Measures the time pressure experienced by participants while performing the task. It considers whether they felt rushed or had adequate time to respond to game requirements.
- *Performance:* Captures participants' satisfaction with how well they believe they performed. This self-assessment provides insight into their confidence in their ability to complete the task effectively.
- *Effort:* Represents the amount of mental or physical work that participants felt they needed to invest beyond the basic task requirements. It indicates how much additional energy or focus was required to manage the workload.
- *Frustration:* Reflects the level of stress, irritation, or annoyance experienced during the task. This dimension sheds light on the emotional challenges participants encountered while navigating the game.

7.3 Ethical Considerations

The ethical integrity of the study was of utmost importance and adhered to the guidelines set by the university's ethical review board. Prior to participation, all individuals were

[1] See NASA-TLX form on: `https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf` (last accessed 04/25)

thoroughly informed about the purpose and structure of the study, as well as the types of data being collected. Importantly, no personal information, such as names or contact details, was recorded during the experiment.

To ensure confidentiality, the data collected were pseudonymized using a codelist method. Each participant was assigned a unique code, and this codelist was kept separately in a secure, hardcopy format stored within the department. The only link between the participants' identities and their corresponding data was this codelist, which was accessible exclusively to authorized personnel. This method ensured that even in the event of a data breach, participants' identities could not be traced back to their data.

In line with data protection regulations, all personal data will be permanently deleted by 31st December 2024, at the latest. This includes the destruction of the codelist and any other identifiers that may connect participants to the study. The pseudonymized research data, however, may be retained for further analysis and publication, as it no longer includes personal identifiers and thus respects participant's privacy.

Each participant was also required to sign a declaration of consent form before taking part in the study. This form was crafted in accordance with the ethical standards of the university and outlined the study's aims, the nature of the tasks involved, and the participants' rights, including their right to withdraw from the study at any time without penalty. This ensured that all participants were aware of their involvement and gave fully informed consent.

Additionally, only individuals over the age of 18 were allowed to participate in the study. This criterion was established because the study aims to determine the influence of timing in an explanation, with potential applications in the design of Autonomous Vehicles. Understanding how quickly drivers typically respond to stimuli is critical for this purpose. Since the minimum legal driving age in Germany (where the study was conducted) is 18, limiting participation to this age group ensures the findings are relevant to the target demographic.

7.4 Data Collection

In this study, data collection focused on capturing essential metrics related to participants' response times and performance, along with basic demographic information. The following types of data were collected:

1. *Reaction Time:*
    - Phase 1 (*Reaction Time Determination*): During this phase, reaction times were recorded as participants responded to colour cues by pressing the corre-

sponding arrow key. This initial phase served to establish a baseline reaction time for each participant.

- Phase 2 (*Reactive Game*): Reaction times were also recorded as participants interacted with the game by collecting coins of the target colour. This allowed for a comparison of reaction times between the controlled settings of Phase 1 and the dynamic, game-based setting of Phase 2.

2. *Game Score:* In Phase 2, participants' scores were recorded based on the number of correctly collected coins that matched the displayed target colour. This score indicated the participant's ability to accurately follow instructions in a time-sensitive, changing environment.

3. *Demographic Information:* Basic demographic data, specifically age and gender, were collected. This information helped identify any performance patterns related to these demographic variables.

4. *User feedback:* After completing the game, the participants were given the NASA-TLX forms to determine the participants' subjective workload.

7.5 Results and Analysis

This section presents the results from both phases of the study, focusing on reaction times, timing of explanations, and participants' subjective workload assessments. The study had 17 participants, with an average age of 44.7 years (SD = 16.4). The gender distribution included 10 female participants, highlighting a fairly balanced sample that allows for general analysis across genders.

Reaction Time Analysis

In the *reaction time determination* phase, participants demonstrated varied response times depending on the direction of the arrow. The slowest response was observed for the down arrow, with an average reaction time of 3.16 s, indicating that pressing the down arrow was either more effortful or less intuitive compared to the other directions. In contrast, the up arrow had the fastest average reaction time of 1.2 s, suggesting it required less cognitive or physical effort, possibly due to a more intuitive response mechanism.

During the *reactive game* phase, across all trials in the experimental round, where participants responded to explanations while collecting target-coloured coins, the average

reaction time across all trials was found to be 2.58 s. This value aligns closely with the predicted 3 s optimal timing predicted by our model in chapter 3, suggesting that participants were able to process and respond to explanations efficiently within this timeframe.

In our study setup, both Salience and Effort were controlled to remain consistent:

- Salience: Explanations were displayed in a uniform format each time, ensuring that participants knew where to look and what to expect visually. This consistency likely contributed to stable reaction times.
- Effort: The explanations appeared in the same location on the screen each time, minimizing any additional cognitive load from seeking out the explanation.

These findings support the model's assumption that a consistent 3 s window allows sufficient time for processing and acting on explanations, particularly when Salience and Effort are kept constant.

User Feedback Analysis

The NASA-TLX data, obtained from the user feedback form, was analysed to assess the perceived workload experienced by participants during the game. Figure 7.5 shows the distribution of NASA-TLX scores across various dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.

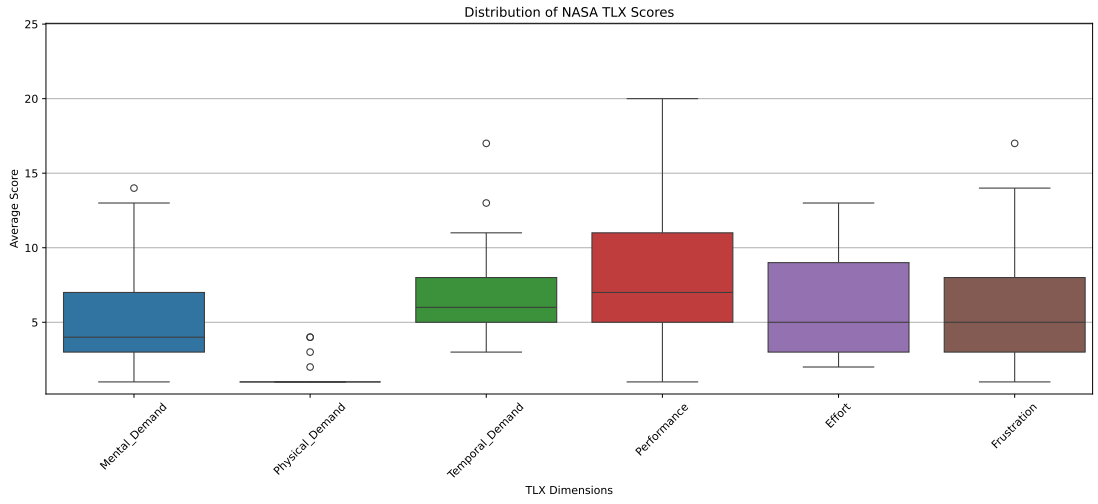


Figure 7.5: Boxplot of NASA-TLX scores across six dimensions, showing medians, and outliers

- *Physical Demand*: Rated the lowest among dimensions, reflecting the minimal physical effort required for the task, consistent with the simple arrow-key interactions.
- *Mental Demand*: Scored slightly higher, indicating that participants found the task cognitively engaging, likely due to the need to focus on colour-matching and directional responses.
- *Temporal Demand and Effort*: Temporal demand and effort dimensions showed moderate scores, reflecting the time-sensitive nature of the task and the sustained attention, which may have contributed to increased cognitive workload.
- *Performance and Frustration*: Scores in these categories varied. While some participants reported satisfaction with their performance, others expressed frustration, often tied to difficulties in achieving target outcomes. Frustration was notably higher than physical demand, highlighting the cognitive challenges associated with meeting the task objectives.

The correlation matrix in fig. 7.6 provides insights into relationships between demographic data and their perceived workload, obtained from NASA-TLX variables. The results can be categorised into significant and marginally significant correlations.

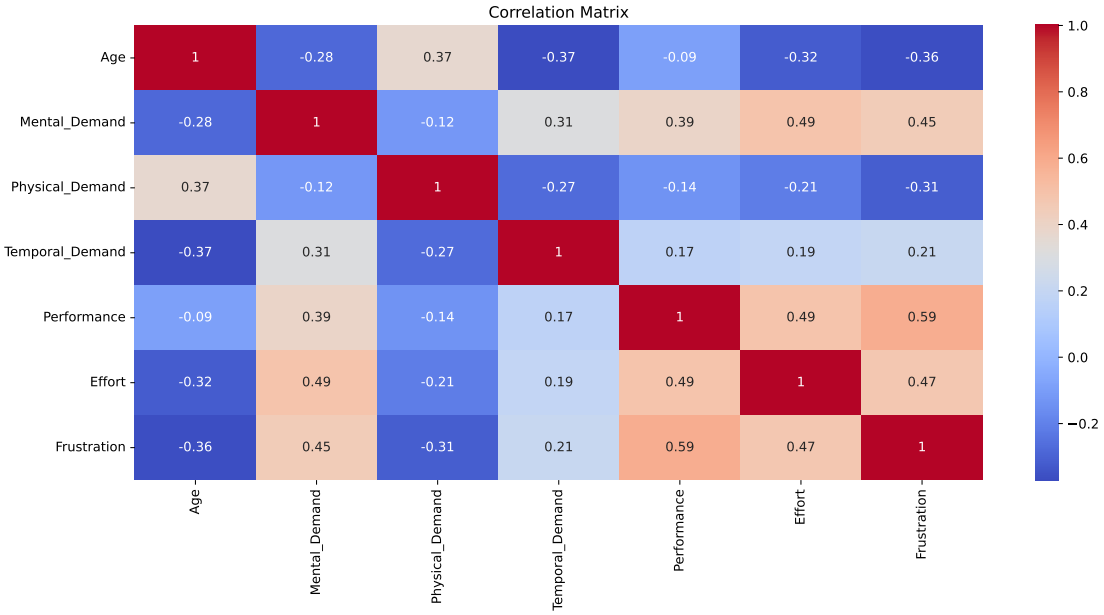


Figure 7.6: Correlation Matrix of NASA-TLX scores and Demographic data

Significant correlations:

- *Mental Demand and Effort (r = 0.49, p = 0.0076)*: This strong positive correlation suggests that as mental demand increases, effort also increases. This indicates that participants who found the task mentally demanding also perceived it as requiring greater effort.
- *Mental Demand and Performance (r = 0.39, p = 0.0169)*: A moderate positive correlation indicates that higher mental demand is associated with higher performance scores. This may indicate that participants exerted more cognitive effort to maintain performance levels when the mental demand was high.
- *Mental Demand and Frustration (r = 0.45, p = 0.0145)*: A strong correlation between mental demand and frustration suggests that when tasks require greater mental effort, they also lead to higher levels of frustration.

Marginally Significant Correlations:

- *Age and Temporal Demand (r = -0.37, p = 0.0797)*: A negative correlation suggests that older participants experience lower temporal demand, but the p-value (0.0797) is slightly above the significance threshold. This could indicate that older individuals take a more measured approach to task completion, potentially due to greater experience in managing cognitive load.
- *Age and Frustration (r = -0.36, p = 0.0795)*: Similar to the above, this negative correlation suggests that older participants report lower frustration levels, but the result is not statistically strong. This could be because older individuals may have better emotional regulation or are less concerned about task performance.
- *Effort and Frustration (r = 0.47, p = 0.0935)*: A moderately strong correlation suggests that higher effort may lead to increased frustration, but the p-value (0.0935) is slightly above the acceptable threshold. This makes sense from a cognitive workload perspective: tasks requiring more effort are often perceived as more frustrating.

While these correlations are not fully significant, they still indicate potential relationships worth exploring in future studies with larger participant groups. These findings offer a detailed view of participant performance and experiences across both study phases. Although the game had low physical demands across all ages, the mental workload and pressure to perform at pace induced a moderate level of cognitive demand.

7.6 Discussion

The results of this study highlight that even a simple, single-word instruction (abbreviated explanation) can serve as a valuable baseline for understanding how early an explanation must be delivered for users to comprehend and act on it in time. This minimalist approach was deliberately chosen not as a solution in itself, but rather as a foundational reference point. In time-critical systems like autonomous vehicles (AVs), knowing the minimal timing window required for comprehension is essential before introducing more complex, context-rich explanations that may require additional cognitive effort to process. This baseline is particularly useful in aligning with the prior chapter result (chapter 3), where explanation timing is emphasised as a key factor in AV user interaction. While this study used simple, uniform instructions to control cognitive load and ensure comparability, future research must investigate richer, more adaptive explanations. These will likely vary in effectiveness across users since individual differences in processing capacity, familiarity with automation, or even situational stress can influence comprehension. Importantly, more context often means more complexity, and striking the right balance between informativeness and mental load will be crucial.

The data collected—reaction times, game scores, and demographic details (age and gender)—were selected to measure task performance and cognitive workload without overwhelming the participants. Reaction time data proved especially informative. For example, the average reaction time of 3.16 s for less intuitive actions (like pressing the down arrow) revealed the relative difficulty in processing certain commands. In contrast, the average reaction time of 2.58 s during explanation trials closely matched the model's predicted optimal timing in chapter 3, reinforcing the relationship between explanation delivery and mental workload. Age was found to correlate moderately with several NASA-TLX dimensions—particularly temporal demand and frustration—though not always at statistically significant levels. Interestingly, mental demand showed strong and statistically significant correlations with both effort ($p = 0.0076$) and frustration ($p = 0.0145$). This suggests that increased cognitive demands reliably lead to greater perceived effort and emotional strain, confirming that explanation complexity has measurable impacts on user workload. These findings underline the importance of carefully managing cognitive load when introducing contextual explanations in real-time AV systems.

Although the sample was relatively diverse in age (mean = 44.7 years, SD = 16.4), and included 10 female participants, the overall sample size of 17 remains a limitation, reducing statistical power and the generalizability of the results. Future studies with

larger and more balanced participant groups are needed to validate these patterns more robustly.

The study was conducted in a university shop situated in a city centre. This setting offered access to a broad participant pool, avoiding the homogeneity often found in lab-based studies. However, the natural distractions in the environment—such as ambient noise and foot traffic—may have introduced variability in performance. Nevertheless, the realistic nature of the environment improved the ecological validity of the findings and demonstrated that participants could still meaningfully engage with explanations despite background disturbances.

A few key limitations should be acknowledged. The gamified setup, while useful for maintaining engagement and controlling conditions, does not fully capture the stakes or distractions of real-world AV scenarios. Furthermore, the use of abbreviated explanations, i.e. one-word instructions—while useful for establishing a baseline—offers limited insight into how users respond to contextual or adaptive explanations, which are critical for more nuanced real-world decision-making. Additionally, relying on self-reported NASA-TLX measures introduces potential bias and subjectivity into workload assessment. The simplicity of both the tasks and the explanations may have underestimated the real cognitive demands of dynamic environments like driving. However, this simplicity also enabled a clearer view of baseline cognitive load and timing thresholds, which future systems can build upon.

Future research should explore how context-aware explanations affect cognitive workload and reaction times. This includes integrating adaptive models that tailor explanation timing and content to the individual user's needs and mental state. Studies in driving simulators or real-world AV settings could provide a more comprehensive understanding of user behaviour under more realistic conditions. Moreover, further research should aim to identify the tipping point at which more context begins to hinder rather than help, particularly in time-critical scenarios. Balancing informativeness with cognitive simplicity will be essential in designing effective human-AI interaction strategies for AVs and beyond.

7.7 Chapter Summary

This study demonstrates that the timing of explanations plays a critical role in user performance and comprehension in interactive environments. Through a game-based user study, we observed that explanation timing significantly impacts both reaction time and cognitive workload. These findings have important implications for the design

of interactive systems, where timing should be considered as a key factor in enhancing user experience and task efficiency.

8 Conclusion

Contents

The conclusion of this thesis is divided into four parts: Summary, limitations, next steps, and directions for future work. In section 8.1, we provide an overview of the key contributions and findings from each chapter of the thesis, followed by the limitations of our work in section 8.2. We discuss the possible research directions that have been partially explored in section 8.3. In section 8.4, we explore potential directions for extending this research in the future.

8.1 Summary

The work of this research stemmed from the lack of research into the precise determination of explanation timing for humans based on their cognitive workload. To address this gap, three key research questions were developed:

1. Developing a Framework for Optimal Atomic Explanation Timing
2. Enhancing the Framework for Multi-Step Explanations
3. Applying the Framework Across Multiple User

These questions guided the iterative design and evaluation of the study, with each chapter of this dissertation addressing a specific aspect of the framework. The following sections outline the progression of this research, showcasing how these questions were explored and answered systematically.

Chapter 3 addressed the first research question by establishing the foundational framework for the timing and delivery of atomic explanations to a single user. This chapter

explored strategies to determine the optimal timing for explanations, considering the user's cognitive workload, as modeled by the SEEV model. The findings of this chapter laid the groundwork for subsequent chapters.

In chapter 4, the framework from chapter 3 was extended to accommodate multi-step explanations, addressing the second research question. This chapter investigated how delivering explanations in multiple steps might influence cognitive workload and explored scenarios where users might request partial explanations versus situations where a single, comprehensive explanation would suffice. These findings highlighted the flexibility required in explanation delivery based on user needs.

Building on these foundations, chapter 5 applied the framework to multi-user contexts, expanding on the concepts introduced in chapter 3. This chapter explored how the timing of atomic explanations could adapt to accommodate the diverse needs of multiple users. It demonstrated the adaptability of the SEEV-reactive game in tailoring explanations based on individual user preferences and attention.

In chapter 6, the frameworks from chapter 4 and chapter 5 were combined to investigate multi-step explanations in multi-user environments. This chapter revealed that variations in user attention could result in some users requesting partial explanations while others required complete ones. These findings underscored the importance of dynamic and context-sensitive explanation strategies for effectively addressing the needs of multiple users simultaneously.

Finally, chapter 7 synthesized the insights from the earlier chapters through a user study conducted in a gamified environment. In this study, participants played a coin-collection game where instructions about the target coin colour were displayed. The results showed that users generally took about 3 s to collect the correct coin after the instruction was presented. These findings validated the results from chapter 3, confirming that the optimal timing for providing explanations is approximately 3 s before the corresponding action is required.

8.2 What's Missing and What Can Be Better?

Every research endeavour, no matter how comprehensive, comes with its limitations, and our work is no exception. Identifying these limitations is crucial for guiding future improvements and extending the applicability of the findings. The major limitations of our work are as follows:

- *Subjectivity in Reward/Cost Values of the SEEV Model:* A significant limitation lies in the reward and cost values used in the SEEV (Salience, Effort, Expectancy,

Value) model, which are central to our research. These values are based on educated guesses, obtained via a trial-and-error method, rather than empirical validation or rigorous parameter estimation. This reliance introduces potential biases and reduces the generalizability of the results across different contexts.

- *Limited Validation in Real-world Scenarios:* The experiment conducted (in chapter 7) primarily relied on controlled conditions. While this setup was essential for isolating variables and testing hypotheses, it may not fully capture the complexity and unpredictability of real-world applications, particularly in dynamic and high-stakes environments.

- *Simplistic Representation of Human Cognition and Behaviour:* Although the SEEV model incorporates key cognitive factors, it simplifies the nuanced interplay between attention, trust, emotions, and decision-making in humans. Real-world cognition often involves factors like fatigue, stress, or social influences, which were not accounted for in our work.

- *Assumptions of Homogeneity Across Users:* The work assumes that all users have similar cognitive abilities and react uniformly to explanations and decision-making scenarios. This assumption may overlook individual differences, such as cognitive load tolerance, prior knowledge, or cultural backgrounds, which can significantly impact the model's effectiveness.

- *Static Parameters in Dynamic Systems:* The SEEV model parameters were treated as static throughout the study. However, real-world scenarios are highly dynamic, with constantly changing conditions and user states. Adapting the model to accommodate such variability remains an unexplored area in this work.

- *Focus on Isolated Scenarios:* The scenarios studied were often isolated and task-specific, rather than interconnected, as they would be in a multi-tasking environment. This limitation may restrict the model's applicability in situations where users must divide attention among several competing tasks.

- *Exclusion of Long-term Effects:* The research primarily focused on short-term outcomes, such as immediate reactions and decision-making. Long-term effects, such as the evolution of trust, attention patterns, or user learning over extended periods, were not addressed.

By acknowledging these limitations, we aim to provide a clearer perspective on the boundaries of our research and encourage future work to address these challenges.

8.3 Next Steps

The work presented in this thesis offers numerous opportunities for further exploration in various directions. Some potential directions, which have been discussed in our prior research, are outlined below:

1. *Dialogue-based explanations:* In our paper [Bai+22], we introduced the idea that explanations are a dynamic process of updating beliefs. We explored the importance of convincing explanations in Autonomous Vehicles and proposed a formal approach for generating justified explanations that take into consideration the addressee's trust and attention. We also discussed how a single explanation might not be sufficient to fully clarify something, emphasizing that explanation is a communication process where the recipient/addressee's beliefs are updated step by step, by the explainer, through the provided explanations.

2. *Correlation between content and timing of explanations:* In this thesis, as outlined in section 1.4, we explored the optimal timing for providing explanations with constant or no content. However, in real-world situations, content plays a crucial role in understanding. Our paper [RHB23] proposes a game-theoretic approach to develop a Constantly Informing System (CIS) that accounts for human factors. Using psychological models of human emotions and cognition, we create a formal Human Model (HM) that interacts with the CIS. Our aim is to optimize the effectiveness of the CIS by determining what content should be communicated to the user and when, based on the user's evolving information base, mood, and abilities.

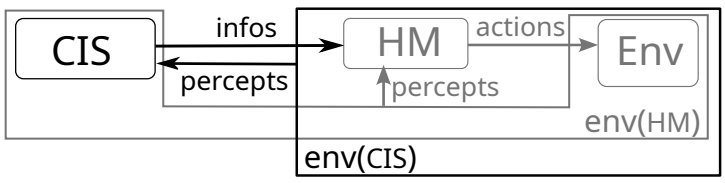


Figure 8.1: Information provision as a game

Figure 8.1 shows how CIS observes its environment and provides information (explanation) to HM. As shown in Figure 8.2, we highlight the importance of integrating emotion generation and cognitive processes in the development of CIS. We propose utilizing psychological models and cognitive architectures to create a HM that interacts with the CIS.

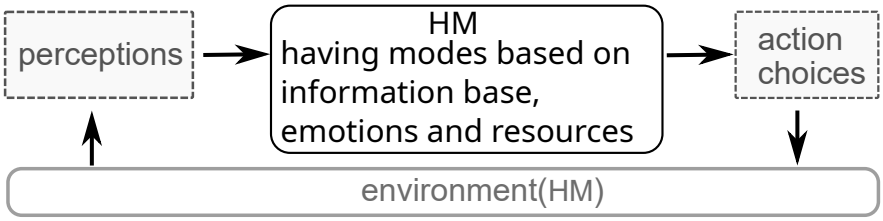


Figure 8.2: HM perceives its environment and chooses actions. During a scenario HM changes its mode, reflecting a change in its cognitive resources, emotions, and information

3. *Conflict resolution using explanations:* Another avenue for further research is leveraging explanations for conflict resolution. This is discussed in our vision paper [BF24b], where we explain how providing (timely) explanations in autonomous systems can enhance trust and reduce frustration in humans.

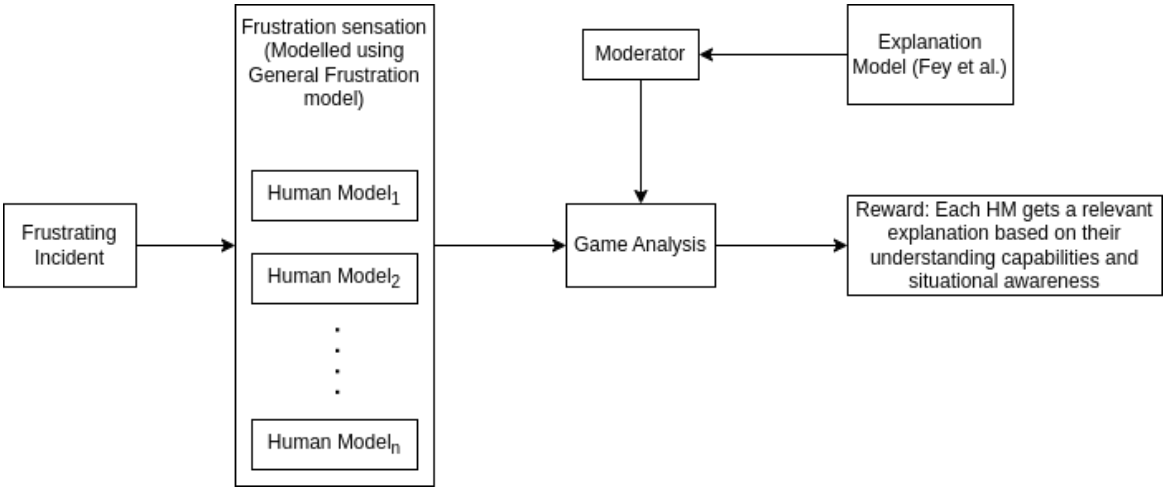


Figure 8.3: Game structure of a game with a moderator, frustration models and an explanation model

In this paper, we propose a framework for an explanation game that involves a moderator to deliver meaningful explanations to humans and reduce frustration by improving their understanding of decision-making. We also suggest a game-theoretic framework, which includes a self-explaining system [FFD22], that provides contextually relevant explanations to humans, with the moderator facilitating explanation delivery to various Human Models (HMs). Figure 8.3 shows the game structure of this explanation game.

4. *Risk mitigation using explanations:* In [Sch+25], we propose the use of explanations in high-risk conflict situations for risk mitigation. Through game-theoretic analysis, we demonstrate how tailored explanations for different participants (Autonomous Traffic Agents (ATAs) and humans) can effectively reduce risks. We advocate for incorporating these explanations into real-time explainability frameworks, particularly when explaining to human users.

   In fig. 8.4 and fig. 8.5, we present two example situations that highlight the importance of explanations as safety measures in various conflict situations.

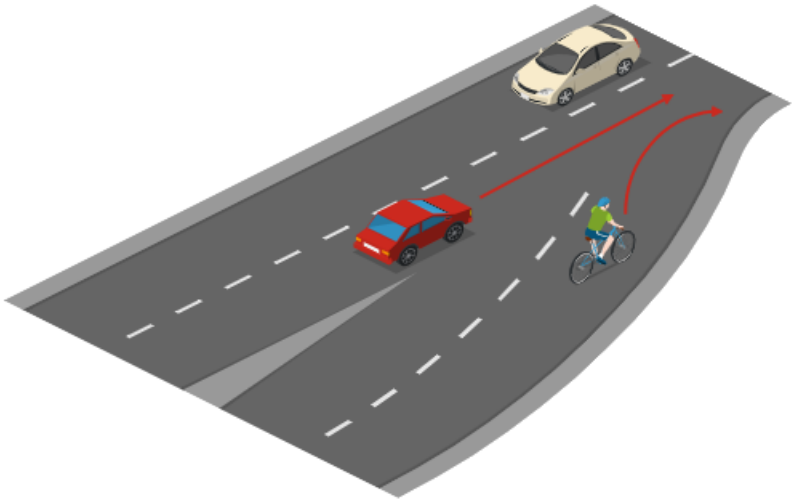


Figure 8.4: *Lane merging*: A bicyclist wants to drive onto the main road from a ramp road.

A conflict scenario involving a vulnerable road user (VRU) is shown in fig. 8.4. In this situation, a bicyclist on a ramp road intends to merge onto the main road ahead of the ego vehicle. The ego vehicle has two high-risk options: (a) maintain its lane and brake, or (b) execute an evasive maneuver into oncoming traffic. Both choices pose significant risks of injury due to potential collisions with the bicyclist or oncoming vehicles, highlighting the need for safety measures.

These risks can be mitigated through explanatory measures, such as discouraging the bicyclist from merging or warning other traffic participants about possible evasive actions.

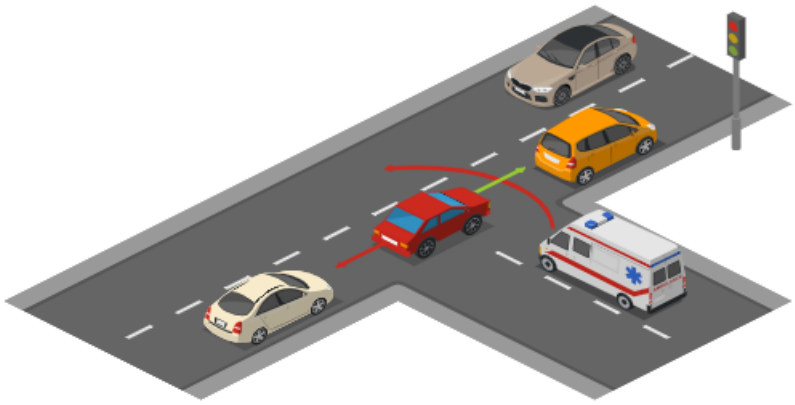


Figure 8.5: *Emergency vehicle*: AV A needs to either reverse or drive forward to let the emergency vehicle pass

Figure 8.5 illustrates a scenario where the ego vehicle is stopped at a traffic light, obstructing an emergency vehicle attempting to enter the intersection and turn left. This situation creates a serious risk of fatal harm due to the blocked path. With vehicles positioned both in front of and behind the ego vehicle, it is impossible for the ego vehicle to yield to the emergency vehicle, leading to a conflict involving multiple agents.

To mitigate this risk, the ego vehicle's manufacturer has two main options: either prevent such scenarios from occurring altogether or equip the ego vehicle with mechanisms to prompt the surrounding vehicles to create space through targeted explanations.

8.4 Directions for Future Work

While we have already presented several ideas that are currently works in progress, there remain numerous opportunities to extend the research introduced in this thesis. Below, we outline some promising directions for future work:

- *Unifying Step-Wise Explanation Frameworks:* Several works focus on step-wise explanations, such as gradual or just-in-time explanations [Bog+20; Ble+23]. A key

challenge is integrating these approaches into a unified framework that accounts for different decision-making contexts and user needs. Future work could investigate models that generalize across different domains, balancing the trade-offs between immediate and delayed explanations while considering long-term learning and trust-building.

- *Multi-Modal and Context-Aware Timing:* In addition to physiological signals, other contextual factors —such as task difficulty, environmental distractions, or social interactions— can influence when an explanation is most effective [Kim+24; MJ13]. Research could explore multi-modal approaches that combine physiological, contextual, and behavioural signals to create more sophisticated, context-aware explanation timing models.

- *Longitudinal Effects of Explanation Timing:* Most existing studies focus on short-term outcomes, such as immediate task performance or user satisfaction [Du+19; KPB18]. However, the long-term impact of explanation timing on trust, learning retention, and mental model development remains underexplored. Future work should investigate how different timing strategies influence user behaviour over extended periods and whether adaptive timing approaches improve cumulative understanding and system acceptance.

- *Emotionally Aware Explanations:* Emotions play a critical role in human decision-making and learning [AP05]. Optimizing the timing of explanations based on a user's emotional state—such as frustration, confidence, or curiosity—could enhance their effectiveness. Emotion-aware systems could leverage sentiment analysis, facial expressions, or tone of voice to adjust explanation delivery dynamically [DMe+05; Gra+13]. Future work could explore emotionally adaptive explanation strategies that provide reassurance during moments of frustration or reduce interruptions when a user is deeply engaged.

- *Ethical and Fairness-Aware Explanation Timing:* The timing of explanations can impact fairness and ethical considerations, particularly in high-stakes applications such as hiring, automotive, healthcare, and finance [Dod+19; SKM22]. Delayed or strategically timed explanations could be misused to manipulate decisions or obscure biases [ZJ23; JZK16]. Future research should focus on developing frameworks that ensure explanation timing promotes fairness, transparency, and informed decision-making while preventing potential manipulation or undue cognitive burden.

Advancing these research directions could lead to more intuitive and human-centred autonomous systems that provide explanations at the right time, ultimately enhancing user trust, learning, and decision-making efficiency.

In conclusion, while the content of an explanation is critical, its timing is equally important. As this thesis has demonstrated, the timing of an explanation significantly influences its effectiveness—reinforcing the principle that *when* an explanation is provided matters just as much as *what* is explained.

List of Figures

List of Tables

Glossary

ACPS Automated Cyber-Physical Systems. iii, v

AI Artificial Intelligence. 2–4, 6, 9

AV Autonomous Vehicle. iii–vi, 1–8, 33, 37–42, 44, 48–50, 55, 58, 60, 69–73, 82, 91, 96, 97, 102, 105, 109, 110

BDD Binary Decision Diagram. 28

BI Backward Induction. 15, 25, 26, 28, 29, 42, 51, 66, 74

CIS Constantly Informing System. 102

CSL Continuous Stochastic Logic. 27, 28

CTMC Continuous-Time Markov Chain. 26, 27

DTMC Discrete-Time Markov Chain. 26, 27

Ef Effort. 32–35, 39, 49, 58, 60, 93

Ex Expectancy. 32–35, 39, 42, 50, 59, 60

FM Fomal Methods. 2, 3

GT Game Theory. 13, 16, 17, 25

HA Human Agent. 7, 8

HCI Human-Computer Interaction. 23

HM Human Model. 102, 103, 110

MAB-EX Manage, Analyze, Build, Explain. 7

MAPE Monitor, Analyze, Plan, Execute. 7

MDP Markov Decision Process. 20, 23, 24, 26, 27, 37, 39, 41, 49–51, 58, 61, 72, 73, 75, 109, 111

ML Machine Learning. 2

NASA-TLX NASA Task Load Index. 85, 90, 92–94, 96, 97, 110

P(A) Probability of Attention. 33, 35, 42, 43, 50, 52, 53

PCTL Probabilistic Computation Tree Logic. 27, 28

PRISM PRISM Model Checker. 15, 26–28

PRISM-games PRISM-games. 27–29

PTA Probabilistic Timed Automata. 26, 28

S Salience. 31, 33–35, 39, 49, 58, 60, 93

SAE Society of Automotive Engineers. 57

SEEV Salience, Effort, Expectancy, Value. iii, v, vi, 15, 31, 33–35, 37, 39, 41, 42, 44, 48–53, 55, 57–61, 64, 67, 69, 71–73, 77, 78, 80, 100, 101, 109

V Value. 32–35, 39, 42, 59

XAI Explainable AI. 9

Bibliography

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

[ABG10] A. Anderson, Daphne Bavelier, and C. Green. "Speed-accuracy tradeoffs in cognitive tasks in action game players." In: *Journal of Vision - J VISION* 10 (Aug. 2010), pp. 748–748. DOI: 10.1167/10.7.748.

[AD13] Allam Hassan Allam and Halina Mohamed Dahlan. "User experience: challenges and opportunities." In: *Journal of information systems research and innovation* 3.1 (2013), pp. 28–36.

[AD94] Rajeev Alur and David L. Dill. "A Theory of Timed Automata." In: *Theor. Comput. Sci.* 126.2 (1994), pp. 183–235. DOI: 10.1016/0304-3975(94)90010-8. URL: https://doi.org/10.1016/0304-3975(94)90010-8.

[AL98] John R. Anderson and Christian Lebiere. *The Atomic Components of Thought.* Lawrence Erlbaum associates, 1998. DOI: 10.4324/9781315805696.

[And96] J. R. Anderson. "Act: a simple theory of complex cognition." In: *American Psychologist* 51 (4 1996), pp. 355–365. DOI: 10.1037/0003-066x.51.4.355.

[AP05] Hyungil Ahn and Rosalind W. Picard. "Affective-Cognitive Learning and Decision Making: A Motivational Reward Framework for Affective Agents." In: *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings.* Ed. by Jianhua Tao, Tieniu Tan, and Rosalind W. Picard. Vol. 3784. Lecture Notes in Computer Science. Springer, 2005, pp. 866–873. DOI: 10.1007/11573548_111. URL: https://doi.org/10.1007/11573548%5C_111.

[Bai+22] Akhila Bairy, Willem Hagemann, Astrid Rakow, and Maike Schwammberger. "Towards Formal Concepts for Explanation Timing and Justifications." In: *30th IEEE International Requirements Engineering Conference Workshops, RE 2022 - Workshops, Melbourne, Australia, August 15-19, 2022.* IEEE, 2022, pp. 98–102. DOI: 10.1109/REW56159.2022.00025. URL: https://doi.org/10.1109/REW56159.2022.00025.

[Bai+25] Akhila Bairy, Mehrnoush Hajnorouzi, Astrid Rakow, Martin Fränzle, and Maike Schwammberger. "Timing Matters - A Study on the Role of Timing in Explanation Delivery." In: *Human Systems Engineering and Design (IHSED2025): Future Trends and Applications* 198 (2025). DOI: http://doi.org/10.54941/ahfe1006782.

[Bai22] Akhila Bairy. "Modeling Explanations in Autonomous Vehicles." In: *Integrated Formal Methods - 17th International Conference, IFM 2022, Lugano, Switzerland, June 7-10, 2022, Proceedings*. Ed. by Maurice H. ter Beek and Rosemary Monahan. Vol. 13274. Lecture Notes in Computer Science. Springer, 2022, pp. 347–351. DOI: 10.1007/978-3-031-07727-2_20. URL: https://doi.org/10.1007/978-3-031-07727-2%5C_20.

[BCM07] Andrea Bunt, Cristina Conati, and Joanna McGrenere. "Supporting interface customization using a mixed-initiative approach." In: *Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI 2007, Honolulu, Hawaii, USA, January 28-31, 2007*. Ed. by David N. Chin, Michelle X. Zhou, Tessa A. Lau, and Angel R. Puerta. ACM, 2007, pp. 92–101. DOI: 10.1145/1216295.1216317. URL: https://doi.org/10.1145/1216295.1216317.

[Bel57] Richard Bellman. *Dynamic Programming*. 1st ed. Princeton, NJ, USA: Princeton University Press, 1957.

[BF23] 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.

[BF24a] 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.

[BF24b] Akhila Bairy and Martin Fränzle. "What if Autonomous Systems had a Game Master? Targeted Explaining with the help of a Supervisory Control System." In: ExEn '24. Lisbon, Portugal: Association for Computing Machinery, 2024, pp. 15–19. ISBN: 9798400705960. DOI: 10.1145/3648505.3648508. URL: https://doi.org/10.1145/3648505.3648508.

[BF25] Akhila Bairy and Martin Fränzle. "Enhancing Multi-user Experience: Optimizing Explanation Timing Through Game Theory." In: *Intelligent Technology for Future Transportation*. Ed. by Abolhassan Razminia and Dinh Hoa Nguyen. Cham: Springer Nature Switzerland, 2025, pp. 106–117. ISBN: 978-3-031-84148-4.

[BFS25] Akhila Bairy, Martin Fränzle, and Maike Schwammberger. *Optimising Timing of Multi-Step Explanations for Multiple Users using Reactive Game*. Accepted to the 7th International Workshop on EXplainable, Trustworthy, and Responsible AI and Multi-Agent Systems (EXTRAAMAS 2025). 2025.

[BIF15] Regina Bernhaupt, Katherine Isbister, and Sara de Freitas. "Introduction to this Special Issue on HCI and Games." In: *Hum. Comput. Interact.* 30.3-4 (2015), pp. 195–201. DOI: 10.1080/07370024.2015.1016573. URL: https://doi.org/10.1080/07370024.2015.1016573.

[Bin+18] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. Ed. by Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox. ACM, 2018, p. 377. DOI: 10.1145/3173574.3173951. URL: https://doi.org/10.1145/3173574.3173951.

[Ble+23] Ignace Bleukx, Jo Devriendt, Emilio Gamba, Bart Bogaerts, and Tias Guns. "Simplifying Step-Wise Explanation Sequences." In: *29th International Conference on Principles and Practice of Constraint Programming, CP 2023, August 27-31, 2023, Toronto, Canada*. Ed. by Roland H. C. Yap. Vol. 280. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023, 11:1–11:20. DOI: 10.4230/LIPICS.CP.2023.11. URL: https://doi.org/10.4230/LIPIcs.CP.2023.11.

[Blu+19] 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, MODELS Companion 2019, Munich, Germany, September 15-20, 2019*. Ed. by Loli Burgueño, Alexander Pretschner, Sebastian Voss, Michel Chaudron, Jörg Kienzle, Markus Völter, Sébastien Gérard, Mansooreh Zahedi, Erwan Bousse, Arend Rensink,

Fiona Polack, Gregor Engels, and Gerti Kappel. IEEE, 2019, pp. 543–548. DOI: 10.1109/MODELS-C.2019.00084. URL: https://doi.org/10.1109/MODELS-C.2019.00084.

[Bog+20] Bart Bogaerts, Emilio Gamba, Jens Claes, and Tias Guns. "Step-Wise Explanations of Constraint Satisfaction Problems." In: *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*. Ed. by Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang. Vol. 325. Frontiers in Artificial Intelligence and Applications. IOS Press, 2020, pp. 640–647. DOI: 10.3233/FAIA200149. URL: https://doi.org/10.3233/FAIA200149.

[Boo+24] Meisam Booshehri, Hendrik Buschmeier, Philipp Cimiano, Stefan Kopp, Jaroslaw Kornowicz, Olesja Lammert, Marco Matarese, Dimitry Mindlin, Amelie Sophie Robrecht, Anna-Lisa Vollmer, Petra Wagner, and Britta Wrede. "Towards a Computational Architecture for Co-Constructive Explainable Systems." In: ExEn '24. Lisbon, Portugal: Association for Computing Machinery, 2024, pp. 20–25. ISBN: 9798400705960. DOI: 10.1145/3648505.3648509. URL: https://doi.org/10.1145/3648505.3648509.

[Car+24] 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.

[Chi09] Michelene T. H. Chi. "Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift." In: 2009. URL: https://api.semanticscholar.org/CorpusID:7682201.

[CLS24] Cheng Chen, Mengqi Liao, and S. Shyam Sundar. "When to Explain? Exploring the Effects of Explanation Timing on User Perceptions and Trust in AI systems." In: *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems, TAS 2024, Austin, TX, USA, Septem-*

ber 16-18, 2024. ACM, 2024, 10:1–10:17. DOI: 10.1145/3686038.3686066. URL: https://doi.org/10.1145/3686038.3686066.

[Det+24] Hannah Deters, Jakob Droste, Anne Hess, Verena Klös, Kurt Schneider, Timo Speith, and Andreas Vogelsang. "The X Factor: On the Relationship between User eXperience and eXplainability." In: *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*. NordiCHI '24. Uppsala, Sweden: Association for Computing Machinery, 2024. ISBN: 9798400709661. DOI: 10.1145/3679318.3685352. URL: https://doi.org/10.1145/3679318.3685352.

[DK17] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML]. URL: https://arxiv.org/abs/1702.08608.

[DMe+05] Sidney D'Mello, Scotty Craig, Barry Gholson, Stan Franklin, Rosalind Picard, and Arthur Graesser. "Integrating affect sensors in an intelligent tutoring system." In: (Jan. 2005). URL: https://api.semanticscholar.org/CorpusID:16331286.

[DO22] Louise A. Dennis and Nir Oren. "Explaining BDI agent behaviour through dialogue." In: *Auton. Agents Multi Agent Syst.* 36.1 (2022), p. 29. DOI: 10.1007/S10458-022-09556-8. URL: https://doi.org/10.1007/s10458-022-09556-8.

[Dod+19] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. "Explaining models: an empirical study of how explanations impact fairness judgment." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 275–285. ISBN: 9781450362726. DOI: 10.1145/3301275.3302310. URL: https://doi.org/10.1145/3301275.3302310.

[DRY23] Na Du, Lionel Robert, and X. Jessie Yang. "Cross-Cultural Investigation of the Effects of Explanations on Drivers' Trust, Preference, and Anxiety in Highly Automated Vehicles." In: *Transportation Research Record* 2677.1 (2023), pp. 554–561. DOI: 10.1177/03611981221100528. eprint: https://doi.org/10.1177/03611981221100528. URL: https://doi.org/10.1177/03611981221100528.

[Du+19] 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.

[El-+19] Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, Rita Sevastjanova, Fabian Sperrle, Thilo Spinner, and Daniel Keim. "Towards XAI: Structuring the Processes of Explanations." In: *Proceedings of the Human-Centered Machine Learning Perspectives Workshop, HCML 2019*. May 2019, 12pp.

[Elb+21] Yusra Elbitar, Michael Schilling, Trung Tin Nguyen, Michael Backes, and Sven Bugiel. "Explanation Beats Context: The Effect of Timing & Rationales on Users' Runtime Permission Decisions." In: *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*. Ed. by Michael D. Bailey and Rachel Greenstadt. USENIX Association, 2021, pp. 785–802. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/elbitar.

[Fer+10] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Christopher Welty. "Building Watson: An Overview of the DeepQA Project." In: *AI Magazine* 31 (Sept. 2010), pp. 59–79. DOI: 10.1609/aimag.v31i3.2303.

[FFD22] Görschwin Fey, Martin Fränzle, and Rolf Drechsler. "Self-Explanation in Systems of Systems." In: *30th IEEE International Requirements Engineering Conference Workshops, RE 2022 - Workshops, Melbourne, Australia, August 15-19, 2022*. IEEE, 2022, pp. 85–91. DOI: 10.1109/REW56159.2022.00023. URL: https://doi.org/10.1109/REW56159.2022.00023.

[Fin+21] Bettina Finzel, David E. Tafler, Stephan Scheele, and Ute Schmid. "Explanation as a Process: User-Centric Construction of Multi-level and Multimodal Explanations." In: *KI 2021: Advances in Artificial Intelligence - 44th German Conference on AI, Virtual Event, September 27 - October 1, 2021, Proceedings*. Ed. by Stefan Edelkamp, Ralf Möller, and Elmar Rueckert. Vol. 12873. Lecture Notes in Computer Science. Springer, 2021, pp. 80–94.

DOI: 10.1007/978-3-030-87626-5_7. URL: https://doi.org/10.1007/978-3-030-87626-5%5C_7.

[FL22] 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.

[FM14] Stephen Flusberg and James Mcclelland. "Connectionism and the emergence of mind." In: Nov. 2014. DOI: 10.1093/oxfordhb/9780199842193.013.5.

[For+11] Vojtech Forejt, Marta Z. Kwiatkowska, Gethin Norman, and David Parker. "Automated Verification Techniques for Probabilistic Systems." In: *Formal Methods for Eternal Networked Software Systems - 11th International School on Formal Methods for the Design of Computer, Communication and Software Systems, SFM 2011, Bertinoro, Italy, June 13-18, 2011. Advanced Lectures*. Ed. by Marco Bernardo and Valérie Issarny. Vol. 6659. Lecture Notes in Computer Science. Springer, 2011, pp. 53–113. DOI: 10.1007/978-3-642-21455-4_3. URL: https://doi.org/10.1007/978-3-642-21455-4%5C_3.

[Fra+14] Stan Franklin, Tamas Madl, Sidney D'Mello, and Javier Snaider. "LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning." In: *IEEE Transactions on Autonomous Mental Development* 6.1 (2014), pp. 19–41. DOI: 10.1109/TAMD.2013.2277589.

[FT91] Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

[GM97] Daniel Granot and Michael Maschler. "The reactive bargaining set: Structure, dynamics and extension to NTU games." In: *Int. J. Game Theory* 26.1 (1997), pp. 75–95. DOI: 10.1007/BF01262514. URL: https://doi.org/10.1007/BF01262514.

[Gra+13] Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. "Automatically Recognizing Facial Indicators of Frustration: A Learning-centric Analysis." In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 159–165. DOI: 10.1109/ACII.2013.33.

[GRE01] Marc Grosjean, David Rosenbaum, and Catherine Elsinger. "Timing and Reaction Time." In: *Journal of experimental psychology. General* 130 (June 2001), pp. 256–72. DOI: 10.1037/0096-3445.130.2.256.

[Gu+20] Yuanyuan Gu, Simeng Gu, Yi Lei, and Hong Li. "From Uncertainty to Anxiety: How Uncertainty Fuels Anxiety in a Process Mediated by Intolerance of Uncertainty." In: *Neural Plasticity* 2020 (Nov. 2020), pp. 1–8. DOI: 10.1155/2020/8866386.

[Gud16] V.N. Gudivada. "Chapter 1 - Cognitive Computing: Concepts, Architectures, Systems, and Applications." In: *Cognitive Computing: Theory and Applications*. Ed. by Venkat N. Gudivada, Vijay V. Raghavan, Venu Govindaraju, and C.R. Rao. Vol. 35. Handbook of Statistics. Elsevier, 2016, pp. 3–38. DOI: https://doi.org/10.1016/bs.host.2016.07.004. URL: https://www.sciencedirect.com/science/article/pii/S0169716116300451.

[Gui22] Riccardo Guidotti. "Counterfactual explanations and how to find them: literature review and benchmarking." In: *Data Mining and Knowledge Discovery* 38 (Apr. 2022), pp. 1–55. DOI: 10.1007/s10618-022-00831-6.

[Gun+21] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. "DARPA's explainable AI (XAI) program: A retrospective." In: *Applied AI Letters* 2.4 (Dec. 2021). DOI: 10.1002/ail2.61. URL: https://doi.org/10.1002/ail2.61.

[Haj+24] Mehrnoush Hajnorouzi, Astrid Rakow, Akhila Bairy, Jan-Patrick Osterloh, and Martin Fränzle. "What Level of Power Should We Give an Automation?" In: *Dependable Computing – EDCC 2024 Workshops*. Ed. by Behrooz Sangchoolie, Rasmus Adler, Richard Hawkins, Philipp Schleiss, Alessia Arteconi, and Adriano Mancini. Cham: Springer Nature Switzerland, 2024, pp. 14–21. ISBN: 978-3-031-56776-6. DOI: 10.1007/978-3-031-56776-6_2.

[Has+18] Jacob Haspiel, Na Du, Jill Meyerson, Lionel P. Robert Jr., Dawn M. Tilbury, X. Jessie Yang, and Anuj K. Pradhan. "Explanations and Expectations: Trust Building in Automated Vehicles." In: *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, Chicago, IL, USA, March 05-08, 2018*. Ed. by Takayuki Kanda, Selma Sabanovic, Guy Hoffman, and Adriana Tapus. ACM, 2018, pp. 119–120. DOI: 10.1145/3173386.3177057. URL: https://doi.org/10.1145/3173386.3177057.

[Hei12] Aviad Heifetz. "Non-credible threats, subgame perfect equilibrium and backward induction." In: *Game Theory: Interactive Strategies in Economics and Management*. Ed. by JudithTranslator Yalon-Fortus. Cambridge University Press, 2012, pp. 317–332.

[Hof+18] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. "Metrics for Explainable AI: Challenges and Prospects." In: *CoRR* abs/1812.04608 (2018). arXiv: 1812.04608. URL: http://arxiv.org/abs/1812.04608.

[How60] Ronald A. Howard. *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press, 1960.

[HS17] Bradley Hayes and Julie A. Shah. "Improving Robot Controller Transparency Through Autonomous Policy Explanation." In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*. Ed. by Bilge Mutlu, Manfred Tscheligi, Astrid Weiss, and James E. Young. ACM, 2017, pp. 303–312. DOI: 10.1145/2909824.3020233. URL: https://doi.org/10.1145/2909824.3020233.

[HS88] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In: *Human Mental Workload*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Advances in Psychology. North-Holland, 1988, pp. 139–183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869.

[HWC06] William J Horrey, Christopher D Wickens, and Kyle P Consalus. "Modeling Drivers' Visual Attention Allocation While Interacting With In-Vehicle Technologies." eng. In: *Journal of experimental psychology. Applied* 12.2 (2006), pp. 67–78. ISSN: 1076-898X. DOI: 10.1037/1076-898X.12.2.67.

[Int21] SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104*. Tech. rep. 2021. URL: https://web.archive.org/web/20211220101755/https://www.sae.org/standards/content/j3016%5C_202104/.

[Joh86] P. N. Johnson-Laird. *Mental models: towards a cognitive science of language, inference, and consciousness*. USA: Harvard University Press, 1986. ISBN: 0674568826.

[JZK16] Samuel G. B. Johnson, Marianna Zhang, and Frank Keil. "Decision-Making and Biases in Causal-Explanatory Reasoning." In: *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Recognizing and Representing Events, CogSci 2016, Philadelphia, PA, USA, August 10-13, 2016*. Ed. by Anna Papafragou, Daniel Grodner, Daniel Mirman, and John C. Trueswell. cognitivesciencesociety.org, 2016. URL: https://mindmodeling.org/cogsci2016/papers/0343/index.html.

[KA97] Douglas S. Krull and Craig A. Anderson. "The Process of Explanation." In: *Current Directions in Psychological Science* 6.1 (1997), pp. 1–5. DOI: 10.1111/1467-8721.ep11512447.

[Kah73] Daniel Kahneman. *Attention and Effort*. Prentice-Hall, 1973.

[Kan00] Barry Kantowitz. "Attention and Mental Workload." In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44 (July 2000), pp. 3–456. DOI: 10.1177/154193120004402121.

[KF19] Abhishek N. Kulkarni and Jie Fu. "Opportunistic Synthesis in Reactive Games under Information Asymmetry." In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, Dec. 2019, pp. 5323–5329. DOI: 10.1109/cdc40024.2019.9029851. URL: http://dx.doi.org/10.1109/CDC40024.2019.9029851.

[Kim+23] Gwangbin Kim, Dohyeon Yeo, Taewoo Jo, Daniela Rus, and Seungjun Kim. "What and When to Explain?: On-road Evaluation of Explanations in Highly Automated Vehicles." In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7.3 (2023), 104:1–104:26. DOI: 10.1145/3610886. URL: https://doi.org/10.1145/3610886.

[Kim+24] Gwangbin Kim, Seokhyun Hwang, Minwoo Seong, Dohyeon Yeo, Daniela Rus, and Seungjun Kim. "TimelyTale: A Multimodal Dataset Approach to Assessing Passengers' Explanation Demands in Highly Automated Vehicles." In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8.3 (2024), 109:1–109:60. DOI: 10.1145/3678544. URL: https://doi.org/10.1145/3678544.

[KNP02] M. Kwiatkowska, G. Norman, and A. Pacheco. "Model Checking CSL Until Formulae with Random Time Bounds." In: *Proc. 2nd Joint International Workshop on Process Algebra and Probabilistic Methods, Performance Mod-*

eling and Verification (PAPM/PROBMIV'02). Ed. by H. Hermanns and R. Segala. Vol. 2399. LNCS. Springer, 2002, pp. 152–168.

[KNP07] Marta Z. Kwiatkowska, Gethin Norman, and David Parker. "Stochastic Model Checking." In: *Formal Methods for Performance Evaluation, 7th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2007, Bertinoro, Italy, May 28-June 2, 2007, Advanced Lectures*. Ed. by Marco Bernardo and Jane Hillston. Vol. 4486. Lecture Notes in Computer Science. Springer, 2007, pp. 220–270. DOI: 10.1007/978-3-540-72522-0_6. URL: https://doi.org/10.1007/978-3-540-72522-0%5C_6.

[KNP11] Marta Z. Kwiatkowska, Gethin Norman, and David Parker. "PRISM 4.0: Verification of Probabilistic Real-Time Systems." In: *Computer Aided Verification - 23rd International Conference, CAV 2011, Snowbird, UT, USA, July 14-20, 2011. Proceedings*. Ed. by Ganesh Gopalakrishnan and Shaz Qadeer. Vol. 6806. Lecture Notes in Computer Science. Springer, 2011, pp. 585–591. DOI: 10.1007/978-3-642-22110-1_47. URL: https://doi.org/10.1007/978-3-642-22110-1%5C_47.

[Koo+15] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. "Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance." In: *International J. on Interactive Design and Manufacturing* 9.4 (2015), pp. 269–275.

[Koo+16] 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.

[KPB18] Moritz Koerber, 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.

[Kru99] R. Krull. "Science, explanation, instruction." In: *IPCC 99. Communication Jazz: Improvising the New International Communication Culture. Proceedings 1999 IEEE International Professional Communication Conference (Cat. No.99CH37023)*. 1999, pp. 315–323. DOI: 10.1109/IPCC.1999.799142.

[Kul+13a] Todd Kulesza, Simone Stumpf, Margaret M. Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. "Too much, too little, or just right? Ways explanations impact end users' mental models." In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing, San Jose, CA, USA, September 15-19, 2013*. Ed. by Caitlin Kelleher, Margaret M. Burnett, and Stefan Sauer. IEEE Computer Society, 2013, pp. 3–10. DOI: 10.1109/VLHCC.2013.6645235. URL: https://doi.org/10.1109/VLHCC.2013.6645235.

[Kul+13b] Todd Kulesza, Simone Stumpf, Margaret M. Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. "Too much, too little, or just right? Ways explanations impact end users' mental models." In: *2013 IEEE Symposium on Visual Languages and Human Centric Computing, San Jose, CA, USA, September 15-19, 2013*. Ed. by Caitlin Kelleher, Margaret M. Burnett, and Stefan Sauer. IEEE Computer Society, 2013, pp. 3–10. DOI: 10.1109/VLHCC.2013.6645235. URL: https://doi.org/10.1109/VLHCC.2013.6645235.

[Kwi+20] M. Kwiatkowska, G. Norman, D. Parker, and G. Santos. "PRISM-games 3.0: Stochastic Game Verification with Concurrency, Equilibria and Time." In: *Proc. 32nd International Conference on Computer Aided Verification (CAV'20)*. Vol. 12225. LNCS. Springer, 2020, pp. 475–487.

[Lan+21] 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.

[Lew86] David Lewis. "Causal Explanation." In: *Philosophical Papers Vol. II*. Ed. by David K. Lewis. Oxford University Press, 1986, pp. 214–240.

[LFV21] Michele Loi, Andrea Ferrario, and Eleonora Viganò. "Transparency as design publicity: explaining and justifying inscrutable algorithms." In: *Ethics Inf. Technol.* 23.3 (2021), pp. 253–263. DOI: 10.1007/S10676-020-09564-W. URL: https://doi.org/10.1007/s10676-020-09564-w.

[Lit94] Michael L. Littman. "Markov games as a framework for multi-agent reinforcement learning." In: *Machine Learning Proceedings 1994*. Ed. by William W. Cohen and Haym Hirsh. San Francisco (CA): Morgan Kaufmann, 1994, pp. 157–163. ISBN: 978-1-55860-335-6. DOI: https://doi.org/10.1016/B978-1-55860-335-6.50027-1. URL: https://www.sciencedirect.com/science/article/pii/B9781558603356500271.

[Lom16] Tania Lombrozo. "Explanatory Preferences Shape Learning and Inference." In: *Trends in Cognitive Sciences* 20.10 (2016), pp. 748–759. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2016.08.001. URL: https://www.sciencedirect.com/science/article/pii/S136466131630105X.

[Mar+17] Anouk van Maris, Hagen Lehmann, Lorenzo Natale, and Beata J. Grzyb. "The Influence of a Robot's Embodiment on Trust: A Longitudinal Study." In: *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*. Ed. by Bilge Mutlu, Manfred Tscheligi, Astrid Weiss, and James E. Young. ACM, 2017, pp. 313–314. DOI: 10.1145/3029798.3038435. URL: https://doi.org/10.1145/3029798.3038435.

[MAT22] MATLAB. *version 9.13.0 (R2022b)*. Natick, Massachusetts: The MathWorks Inc., 2022.

[MAT23] MATLAB. *version 9.14.0 (R2023a)*. Natick, Massachusetts: The MathWorks Inc., 2023.

[MAT24] MATLAB. *version 24.2.0.2712019 (R2024b)*. Natick, Massachusetts: The MathWorks Inc., 2024.

[Mil19] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial Intelligence* 267 (2019), pp. 1–38. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[Min+24] Dimitry Mindlin, Fabian Beer, Leonie Nora Sieger, Stefan Heindorf, Philipp Cimiano, Elena Esposito, and Axel-Cyrille Ngonga Ngomo. "Beyond One-Shot Explanations: A Systematic Literature Review of Dialogue-Based XAI Approaches." In: (Feb. 2024). DOI: 10.21203/rs.3.rs-3998994/v1.

[MJ13] Yashar Moshfeghi and Joemon M. Jose. "An effective implicit relevance feedback technique using affective, physiological and behavioural features." In: *The 36th International ACM SIGIR conference on research and develop-*

ment in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013. Ed. by Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai. ACM, 2013, pp. 133–142. DOI: 10.1145/2484028.2484074. URL: https://doi.org/10.1145/2484028.2484074.

[MKR21] 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.

[MN19] Alex McAvoy and Martin A. Nowak. "Reactive learning strategies for iterated games." In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2223 (Mar. 2019), p. 20180819. ISSN: 1471-2946. DOI: 10.1098/rspa.2018.0819. URL: http://dx.doi.org/10.1098/rspa.2018.0819.

[Nas50] John F. Nash. "Equilibrium points in n-person games." In: *Proceedings of the National Academy of Sciences* 36.1 (1950), pp. 48–49. DOI: 10.1073/pnas.36.1.48. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.36.1.48. URL: https://www.pnas.org/doi/abs/10.1073/pnas.36.1.48.

[Niv07] Eric Nivel. "Ikon flux 2.0." In: 2007. URL: https://api.semanticscholar.org/CorpusID:60599731.

[NM07] 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://press.princeton.edu/titles/7802.html.

[NPS13] Gethin Norman, David Parker, and Jeremy Sproston. "Model Checking for Probabilistic Timed Automata." In: *Formal Methods in System Design* 43.2 (2013), pp. 164–190.

[OR94] Martin J. Osborne and Ariel Rubinstein. *A course in game theory*. electronic edition. Cambridge, USA: The MIT Press, 1994. ISBN: 0-262-65040-1.

[Pap+23] Guglielmo Papagni, Jesse de Pagter, Setareh Zafari, Michael Filzmoser, and Sabine T. Köszegi. "Artificial agents' explainability to support trust: considerations on timing and context." In: *AI Soc.* 38.2 (2023), pp. 947–960. DOI:

10.1007/S00146-022-01462-7. URL: https://doi.org/10.1007/s00146-022-01462-7.

[Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. USA: Cambridge University Press, 2009. ISBN: 052189560X.

[Pia05] Jean Piaget. *The psychology of intelligence*. Routledge, 2005.

[PM18] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. 1st. USA: Basic Books, Inc., 2018. ISBN: 046509760X. DOI: 10.1007/S00146-020-00971-7. URL: https://doi.org/10.1007/s00146-020-00971-7.

[Put94] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN: 978-0-47161977-2. DOI: 10.1002/9780470316887. URL: https://doi.org/10.1002/9780470316887.

[RHB23] 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.

[Rie+23] Martin Riemer, Johanna Bogon, Nele Rußwinkel, Niels Henze, Eva Wiese, David Halbhuber, and Roland Thomaschke. "Time and Timing in Human-Computer Interaction." In: *Mensch und Computer 2023 - Workshopband, Rapperswil, Switzerland, September 3-6, 2023*. Gesellschaft für Informatik e.V., 2023. DOI: 10.18420/MUC2023-MCI-WS05-106. URL: https://doi.org/10.18420/muc2023-mci-ws05-106.

[Riv+21] Florian (4ian) Rivial, Victor Levasseur, Aurélien Vivet, Arthur Pacaud, Franco Maciel, and Todor Imreorov. *GDevelop 5*. Accessed: 2024-05-17. 2021. URL: https://editor.gdevelop.io/.

[Ros+20] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, Michael L. Walters, and Patrick Holthaus. "Evaluating People's Perceptions of Trust in a Robot in a Repeated Interactions Study." In: *Social Robotics - 12th International Conference, ICSR 2020, Golden, CO, USA, November 14-18, 2020, Proceedings*. Ed. by Alan R. Wagner, David Feil-Seifer, Kerstin So-

phie Haring, Silvia Rossi, Thomas Emrys Williams, Hongsheng He, and Shuzhi Sam Ge. Vol. 12483. Lecture Notes in Computer Science. Springer, 2020, pp. 453–465. DOI: 10.1007/978-3-030-62056-1_38. URL: https://doi.org/10.1007/978-3-030-62056-1%5C_38.

[Ros13] Paul Rosenbloom. "The Sigma cognitive architecture and system." In: *AISB Quarterly* 136 (Jan. 2013), pp. 4–13.

[RSG16] 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.

[RTC18] 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. URL: https://doi.org/10.3390/mti2040062.

[Rus16] John Rust. "Dynamic Programming." In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, 2016, pp. 1–26. ISBN: 978-1-349-95121-5. DOI: 10.1057/978-1-349-95121-5_1932-1. URL: https://doi.org/10.1057/978-1-349-95121-5_1932-1.

[RZG14] Anna N. Rafferty, Matei Zaharia, and Thomas L. Griffiths. "Optimally designing games for behavioural research." English (US). In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 470.2167 (July 2014). Copyright: Copyright 2014 Elsevier B.V., All rights reserved. ISSN: 1364-5021. DOI: 10.1098/rspa.2013.0828.

[Sch+19] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. "I can do better than your AI: expertise and explanations." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*. Ed. by Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary. ACM, 2019, pp. 240–251. DOI: 10.1145/3301275.3302308. URL: https://doi.org/10.1145/3301275.3302308.

[Sch+20] Jan Maarten Schraagen, Pia Elsasser, Hanna Fricke, Marleen Hof, and Fabyen Ragalmuto. "Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models." In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64.1 (2020), pp. 339–343. DOI: 10.1177/1071181320641077. eprint: https://doi.org/10.1177/1071181320641077. URL: https://doi.org/10.1177/1071181320641077.

[Sch+21] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R. S. Gerlicher. "ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving." In: *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. Ed. by Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker. ACM, 2021, 161:1–161:12. DOI: 10.1145/3411764.3446647. URL: https://doi.org/10.1145/3411764.3446647.

[Sch+25] Maike Schwammberger, Astrid Rakow, Lina Putze, and Akhila Bairy. "Explain it for Safety: Explanations for Risk Mitigation." In: *Design and Verification of Cyber-Physical Systems: From Theory to Applications*. Ed. by Andreas Rauh, Bernd Finkbeiner, and Paul Kröger. to be published. 2025.

[Sha53] L. S. Shapley. "Stochastic Games*." In: *Proceedings of the National Academy of Sciences* 39.10 (1953), pp. 1095–1100. DOI: 10.1073/pnas.39.10.1095. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.39.10.1095. URL: https://www.pnas.org/doi/abs/10.1073/pnas.39.10.1095.

[She+20] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, Eileen Yang, Xilun Jin, Yuliang Fan, and Katie Driggs Campbell. "To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles." In: *CoRR* abs/2006.11684 (2020). arXiv: 2006.11684. URL: https://arxiv.org/abs/2006.11684.

[Sin06] David Sinreich. "An architectural blueprint for autonomic computing." In: *IBM Autonomic Computing – White Paper* (2006).

[SK22] 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), FMAS/ASYDE@SEFM 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.

[SKM22] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. ""There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1616–1628. ISBN: 9781450393522. DOI: 10.1145/3531146.3533218. URL: https://doi.org/10.1145/3531146.3533218.

[SN71] Herbert A. Simon and Allen Newell. "Human problem solving: The state of the theory in 1970." In: *American Psychologist* 26.2 (1971), p. 145.

[Sta12] Kathleen Stansberry. "One-step, two-step, or multi-step flow: The role of influencers in information processing and dissemination in online, interest-based publics." PhD thesis. University of Oregon, USA, 2012. URL: https://scholarsbank.uoregon.edu/items/5ccb436f-3f26-4aba-9494-6e3b7cb10aca.

[SV15] Eilon Solan and Nicolas Vieille. "Stochastic games." In: *Proceedings of the National Academy of Sciences* 112.45 (2015), pp. 13743–13746. DOI: 10.1073/pnas.1513508112. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1513508112. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1513508112.

[Swe88] 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.

[Taa13] Niels Taatgen. "The Nature and Transfer of Cognitive Skills." In: *Psychological review* 120 (June 2013), pp. 439–471. DOI: 10.1037/a0033138.

[THK24] Sule Tekkesinoglu, Azra Habibovic, and Lars Kunze. *Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap*. 2024. arXiv: 2404.00019 [cs.HC]. URL: https://arxiv.org/abs/2404.00019.

[TK74] A Tversky and D Kahneman. "Judgment under Uncertainty: Heuristics and Biases." In: *Science* 185.4157 (Sept. 1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124. URL: https://www.ncbi.nlm.nih.gov/pubmed/17835457.

[Vis06] Willemien Visser. *The cognitive artifacts of designing*. CRC Press, 2006, p. 280. ISBN: 9780429179693. DOI: 10.1201/9781482269529.

[Wal04] Douglas Walton. "A new dialectical theory of explanation." In: *Philosophical Explorations* 7.1 (2004), pp. 71–89. DOI: 10.1080/1386979032000186863. eprint: https://doi.org/10.1080/1386979032000186863. URL: https://doi.org/10.1080/1386979032000186863.

[WBK23] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. "Categorical and Continuous Features in Counterfactual Explanations of AI Systems." In: *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI 2023, Sydney, NSW, Australia, March 27-31, 2023*. ACM, 2023, pp. 171–187. DOI: 10.1145/3581641.3584090. URL: https://doi.org/10.1145/3581641.3584090.

[Wic+01] Christopher Wickens, John Helleberg, Juliana Goh, Xidong Xu, and William Horrey. "Pilot Task Management: Testing an Attentional Expected Value Model of Visual Scanning." In: *Savoy, IL, UIUC Institute of Aviation Technical Report* (2001).

[Wic15] Christopher Wickens. "Noticing events in the visual workplace: The SEEV and NSEEV models." In: *The Cambridge Handbook of Applied Perception Research*. Cambridge Handbooks in Psychology. Cambridge University Press, 2015, pp. 749–768. DOI: 10.1017/CBO9780511973017.046.

[Wie+19] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. "I Drive - You Trust: Explaining Driving Behavior Of Autonomous Cars." In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. Ed. by Regan L. Mandryk, Stephen A. Brewster, Mark Hancock, Geraldine Fitzpatrick, Anna L. Cox, Vassilis Kostakos, and Mark Perry. ACM, 2019. DOI: 10.1145/3290607.3312817. URL: https://doi.org/10.1145/3290607.3312817.

[WKB22] Greta Warren, Mark T. Keane, and Ruth M. J. Byrne. "Features of Explainability: How users understand counterfactual and causal explanations for

categorical and continuous features in XAI." In: *Proceedings of the Workshop on Cognitive Aspects of Knowledge Representation co-located with the 31st international join conference on artificial intelligence (IJCAI-ECAI 2022), Vienna, Austria, July 23, 2022*. Ed. by Jesse Heyninck, Thomas Meyer, Marco Ragni, Matthias Thimm, and Gabriele Kern-Isberner. Vol. 3251. CEUR Workshop Proceedings. CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3251/paper1.pdf.

[WLB13] Bertram Wortelen, Andreas Lüdtke, and Martin Baumann. "Integrated Simulation of Attention Distribution and Driving Behavior." In: *Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling and Simulation*. Mar. 2013. URL: https://cc.ist.psu.edu/BRIMS2013/archives/2013/BRIMS2013-117.pdf.

[WMR17] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." In: *CoRR* abs/1711.00399 (2017). arXiv: 1711.00399. URL: http://arxiv.org/abs/1711.00399.

[Wor14] Bertram Wortelen. "Das Adaptive-Information-Expectancy-Modell zur Aufmerksamkeitssimulation eines kognitiven Fahrermodells." PhD thesis. University of Oldenburg, Germany, 2014. URL: http://oops.uni-oldenburg.de/1970.

[WS21] Richard Warner and Robert H. Sloan. "Making Artificial Intelligence Transparent: Fairness and the Problem of Proxy Variables." In: *Criminal Justice Ethics* 40.1 (2021), pp. 23–39. DOI: 10.1080/0731129X.2021.1893932. eprint: https://doi.org/10.1080/0731129X.2021.1893932. URL: https://doi.org/10.1080/0731129X.2021.1893932.

[Xu+21] 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.

[Xu+23] Yifan Xu, Joe Collenette, Louise A. Dennis, and Clare Dixon. "Dialogue Explanations for Rule-Based AI Systems." In: *Explainable and Transparent AI and Multi-Agent Systems - 5th International Workshop, EXTRAAMAS 2023, London, UK, May 29, 2023, Revised Selected Papers*. Ed. by Davide Calvaresi, Amro Najjar, Andrea Omicini, Reyhan Aydogan, Rachele Carli, Giovanni Ciatto, Yazan Mualla, and Kary Främling. Vol. 14127. Lecture

Notes in Computer Science. Springer, 2023, pp. 59–77. DOI: 10.1007/978-3-031-40878-6_4. URL: https://doi.org/10.1007/978-3-031-40878-6%5C_4.

[Xu22] Yifan Xu. "Dialogue Explanation With Reasoning for AI." In: *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*. Ed. by Vincent Conitzer, John Tasioulas, Matthias Scheutz, Ryan Calo, Martina Mara, and Annette Zimmermann. ACM, 2022, p. 918. DOI: 10.1145/3514094.3539522. URL: https://doi.org/10.1145/3514094.3539522.

[ZJ23] Joyce Zhou and Thorsten Joachims. "How to Explain and Justify Almost Any Decision: Potential Pitfalls for Accountability in AI Decision-Making." In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 12–21. ISBN: 9798400701924. DOI: 10.1145/3593013.3593972. URL: https://doi.org/10.1145/3593013.3593972.

[ZKG21] Florian Ziesche, Verena Klös, and Sabine Glesner. "Anomaly Detection and Classification to enable Self-Explainability of Autonomous Systems." In: *Design, Automation & Test in Europe Conference & Exhibition, DATE 2021, Grenoble, France, February 1-5, 2021*. IEEE, 2021, pp. 1304–1309. DOI: 10.23919/DATE51398.2021.9474232. URL: https://doi.org/10.23919/DATE51398.2021.9474232.

[ZYR21] Qiaoning Zhang, X. Jessie Yang, and Lionel P. Robert. "What and When to Explain? A Survey of the Impact of Explanation on Attitudes Toward Adopting Automated Vehicles." In: *IEEE Access* 9 (2021), pp. 159533–159540. DOI: 10.1109/ACCESS.2021.3130489. URL: https://doi.org/10.1109/ACCESS.2021.3130489.