Advances in Human Factors of Transportation, Vol. 148, 2024, 307–315
https://doi.org/10.54941/ahfe1005221

Efficiently Explained: Leveraging the SEEV Cognitive Model for Optimal Explanation Delivery

Akhila Bairy and Martin Fränzle

Research Group Foundations and Applications of Systems of Cyber-Physical Systems, Department of Computing Science, Carl von Ossietzky Universität Oldenburg, Germany

ABSTRACT

It is inherent to autonomous systems that they exhibit very complex behaviour and that these complex and flexible patterns of behaviour are in general less comprehensible and foreseeable to humans interacting with the systems. It is generally accepted wisdom that suitable explanations can help humans to understand the functioning of these systems. This, in turn, enhances safety, trust, and societal acceptance through meaningful interaction. Our algorithmic approach starts from the observation that the design of explanations has two essential dimensions to it, namely, content on the one hand and frequency and timing on the other. While there has been extensive research on the substance of explanations, there has been comparatively limited exploration into the precise timing details of explanations. Existing studies on explanation timing often focus on broad distinctions, such as delivering explanations before, during, or after the use of the system. Regarding Autonomous Vehicles (AVs), studies indicate that occupants generally prefer receiving an explanation prior to the occurrence of an autonomous action. However, extended exposure and use of a specific AV may likely diminish the necessity for explanations. Since understanding the explanations can add to (cognitive/mental) workload, this observation suggests the importance of optimising both the frequency —skipping explanations when unnecessary to minimise workload—and the precise timing of explanations, delivering them when they offer the maximum reduction in workload. The interesting fact here is that additional mental workload for the passengers can be caused both by providing and by skipping an explanation: Any explanation that is presented requires cognitive processing for its comprehension, even when its content is considered redundant by the addressee (e.g. due to the explanation content already being familiar to the passenger) or is not memorised (e.g. when an early explanation becomes superimposed by successive events due to the limited capacity of working memory). In contrast, a skipped explanation may prompt the passenger to actively scan the environment for potential cues (e.g. to understand the reasons for an unfamiliar action of the AV) and such an attention strategy induces cognitive workload itself. Concerning the latter effect, Kantowitz has investigated the relation between attention and mental workload and concluded that even simple models of attention are sufficient to predict the mental workload. In this work, we develop a probabilistic reactive game model of mental workload and the impact of explanations on it. It consists of a workload model based on SEEV as a probabilistic component modelling the human and the self-explaining AV function as the other player. The resulting 1.5-player game or Markov Decision Process facilitates to automatically synthesize a rational reactive strategy which will present explanations to the human only when beneficial and then at the optimal time, thereby minimising the cognitive workload of the human.

Keywords: Autonomous vehicles, Explanation timing, Reactive game theory, Attention model, Human-machine-interaction

INTRODUCTION

In an era dominated by highly automated and autonomous systems, the technological landscape is evolving at an unprecedented pace. These advanced systems, characterized by intricate and potentially impactful behaviors, surpass the capabilities of earlier device generations. As humans increasingly interact with these complex and adaptable technologies, the challenge arises in comprehending and predicting their actions. This dynamic presents an opportunity for exploration, as conventional wisdom hints at the importance of offering clear explanations to improve understanding. Such clarity not only fosters safe interaction but also nurtures trust and societal acceptance in the ever-evolving realm of automated systems. This paper focuses on optimizing explanations in autonomous vehicles (AVs) through a specialized algorithm in a game setting. We recognize two key dimensions in explanation formulation: content, representing the substance of the explanation answering the questions of what is happening and why is it happening; and frequency and timing, encapsulating the temporal aspects of explanation delivery. While existing research has made significant progress in understanding the content dimension, there exists a noticeable gap concerning the nuanced timing intricacies associated with delivering explanations. The existing body of research on explanation timing primarily gravitates towards a broad categorization, distinguishing between the presentation of explanations before a use of the system, during its use, or after. In the case of AVs, numerous studies indicate a preference among occupants for receiving explanations prior to the execution of autonomous actions (Du et al., 2019), (Koo et al., 2016), (Ruijten et al., 2018). It does however seem probable that familiarisation due to prolonged exposition to and use of a particular AV will reduce the need for explanation. The comprehension of explanations introduces an inherent workload, necessitating a delicate balance in optimizing both the frequency and precise timing of explanations. This optimization involves the strategic skipping of explanations when deemed unnecessary to mitigate overall workload. Furthermore, it implies that explanations should be timed to offer maximal workload reduction, aligning with the observed preference for pre-action explanations.

The interesting fact here is that additional mental workload for the passengers can be caused both by providing as well as by skipping an explanation. Any explanation that is presented requires cognitive processing for its comprehension, even when its content is considered redundant or not memorised by the addressee. The former may, e.g., occur due to the explanation content already being familiar to the passenger, while the latter may be induced by, e.g., an early explanation becoming superimposed by successive events due to the limited capacity of working memory (cf. Baddeley and Hitch, 1974). Vice versa, a skipped explanation may prompt the passenger to themselves actively scan the environment for potential cues as necessary, e.g., for understanding the reasons for an unfamiliar action of the AV. Such an attention strategy obviously induces a significant cognitive workload itself. Concerning the latter effect, Kantowitz (Kantowitz, 2000) delved into the correlation between attention and mental workload, arriving at the conclusion that even simplistic attention models prove adequate in predicting mental workload.

Since its inception in 1944, mathematical game theory (von Neumann & Morgenstern, 1944) has been a versatile framework for understanding human decision-making across various domains. Despite recent stateful models in neuropsychology and cognitive psychology, the exploitation of stateful game theory in human-machine interaction design is in its early stages. In our work, we are using game theory along with a stateful model of attention called SEEV, developed by Wickens and others (Wickens et al., 2001). By adopting SEEV, we aim to capture the evolving nature of attention over time, providing a more comprehensive framework for assessing and predicting the cognitive demands associated with explanation delivery in the context of AVs.

In the next section we examine the intricate relationship between timing and explanation through an example scenario. We then delve into the SEEV model, discussing its conceptual framework and theoretical foundations, followed by its implementation in a reactive game. The penultimate section focuses on the game results, where we analyze and interpret findings. Our paper concludes by summarizing key contributions and future prospects for our project.

EXAMPLE SCENARIO

To demonstrate the effect of the timing of an explanation on human attention, let us consider the following example shown in (Bairy et al., 2022):

> *At an intersection, an autonomous vehicle* ***v*** *that plans to take a left turn stops despite a green traffic light permitting an uninterrupted left turn.* ***v*** *is prepared to explain to its occupants that it stops to give way to an emergency vehicle, but has to decide whether and when to provide the explanation.*

If the explanation is provided too early in the example scenario, e.g. before it is clear to the passenger that an intersection is ahead and a left turn imminent, the information might be disregarded, with the cognitive workload induced by its processing being wasted. If the explanation is provided too late — even just slightly late — or not provided at all, the occupants will be prone to start their own attention strategy screening the environment to come up with an explanation for the unexpected stopping at a green light, thus increasing the cognitive workload by pursuing active attention. Minimisation of cognitive workload thus critically hinges on the fine-granular optimisation of explanation timing.

ATTENTION MODEL - SEEV

Wickens and others developed a model called SEEV to quantitatively evaluate and predict the attention level of a human in any given situation across different areas of interest (Wickens et al., 2001). The SEEV model was initially developed to predict the attention of a pilot in a cockpit. Later on, Horrey (Horrey et al., 2006) as well as Wortelen (Wortelen, 2014) utilised the SEEV model and derivatives thereof to predict the attention of a driver in road traffic.

SEEV is an acronym for its four attributes representing Salience (S), Effort (Ef), Expectancy (Ex), and Value (V) of an information item or area of interest. *Salience* describes how salient fresh information of the particular type, if becoming available, would be to the human. *Effort* refers to the amount of (physical) effort applied by the human to perceive this new information. *Expectancy* refers to the frequency of new information becoming available and consequently is a dynamic variable describing the expected remaining time to arrival of updated information on the particular item. *Value* is the gain the human expects from the information item. The formula for calculating the probability of attention *P(A)* to an item using SEEV is

$$P(A) = S - Ef + Ex \cdot V \tag{1}$$

Since salience and effort are both related to physical properties of the environment, they are grouped together and referred to as "bottom-up" factors which affect attention (Wickens, 2015). The other two attributes, namely, expectancy and value, are termed "top-down" factors. Note that some of these factors change dynamically, inducing a temporal dynamic of attention that we will exploit for optimising explanation timing. This obviously applies to the top-down factors of expectancy, i.e. the remaining time to arrival of fresh information, and of value, as the value may change by moving into a differently structured environment (e.g., from urban driving to an expressway) or as existing information ages and gets invalidated by the dynamic evolution of environmental states.

THE SEEV MODEL IN A REACTIVE DECISION GAME

As sketched in the introduction, we want to employ the SEEV model as a means for rationally deciding on whether and when to provide an explanation. We, therefore, build a reactive game graph from the SEEV model and synthesise a strategy minimising the expected workload for the human, where the strategic options at each time instant are to either present an explanation or to refrain from doing so currently. As the SEEV model describes autonomous stochastic dynamics, namely a probabilistic model of the human paying attention as a function of the evolution of time and the last time of an explanation, the corresponding game model would be a Markov Decision Process (MDP), a.k.a. 1.5-player game, due to (Howard, 1960). Such a game features a strategic (or 'full') player following a designated strategy against a random (or 'half') player selecting actions at random according to given probability distributions. In our setting, the random player is given by the SEEV model, which decides at random — though with history-dependent probability — when to pay attention to an item, thereby inducing the workload associated with the attention strategy. Our strategic player is the explanation mechanism, which decides strategically when to present an explanation in order to minimise the expected cognitive workload on the human, as suggested in (Bairy et al., 2022). Note that presenting an explanation also induces workload on the human connected to explanation reception and interpretation, albeit generally at lower workload levels than pursuing an active attentive strategy. In our game, we work only on finding the optimal time

for an explanation shown in section Example Scenario. Since there is only a fixed area of interest in this scenario, 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 both Salience and Effort and consequently also their difference can be replaced by a constant in eq. 1. The probability of attention can now be calculated using just the top-down factors as

$$P(A) = Ex \cdot V + c \tag{2}$$

The above SEEV model has been implemented as a dynamic factor impacting workload within a Markov decision process where the decisions concern explanation presentation, and we have employed MATLAB (MATLAB, 2022) to compute the reactive presentation strategy for the explanation. The SEEV game starts n time seconds before the scenario occurs, and ends when the scenario finalises.

At each time step, which is one second, the strategic player, i.e. the explanation presentation machinery, can perform an action that takes it to one of the following three states, namely: *no_expl* indicating the absence of an explanation; *expl* denoting the provision of an explanation; and *no_expl_needed* indicating instances where an explanation is deemed unnecessary. Each of the *{states, actions}* pair of the strategic player is associated with certain costs/rewards. The attention level of the occupant is then assessed using eq. 2, which considers the chosen action by the strategic player and the associated cost.

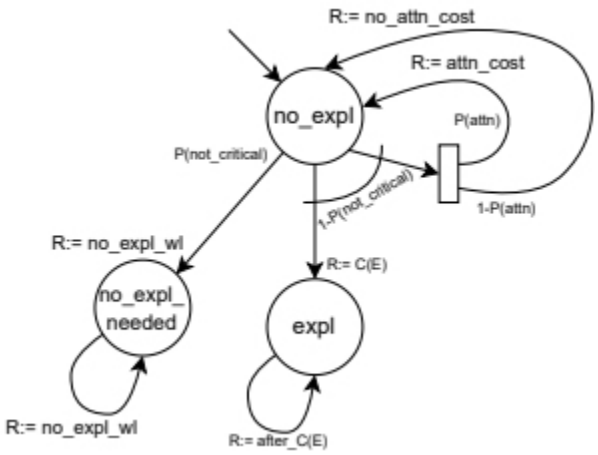

Figure 1: State diagram of the strategic player.

Figure 1 shows a state diagram with the different states and their transitions for the strategic player. The expectancy of the random player builds up with a constant value (exp) over time. this is depicted in the Figure 2.

The value of the SEEV model can also be considered a constant since our game is applied only to the example in the Example Scenario section which has only one area of interest. Costs/rewards are provided for various actions

taken by the strategic player. The rewards depend on the probability of attention (P*(attn)*) and the probability of no critical scenario (P*(not_critical)*). P*(attn)* is calculated using the eq. 2. In the next section we discuss more about how the rewards are calculated.

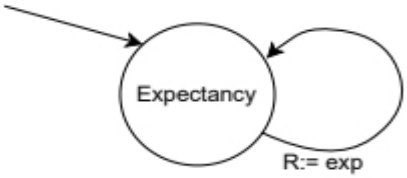

Figure 2: Expectancy of the random player (Baity et al., 2023).

EXPERIMENTAL RESULTS

The goals of the SEEV game were to identify the time when it is ideal to provide the explanation in order to obtain the minimum cognitive workload and to determine the minimum, across all presentation strategies possible, the expected cognitive workload on the human. To determine these factors, rewards need to be assigned for the various state transitions. This is given in the Table 1.

Table 1. MDP rewards.

S	S'	Probability	R
no_expl	no_expl	P(not_critical). P(attn)	0.4
no_expl	no_expl	P(not_critical). (1-P(attn))	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

As this is a finite horizon model, backward Bellman induction is used to calculate the minimum workload that is induced by the attention strategy at any given point in time. The formula to calculate the minimum workload (*min_wl*) is given by eq. 3. At certain times, the explanation might not be required if the user has already evaluated the surroundings or the situation resolves itself. This is taken into consideration as a probability of no critical scenario (P*(not_critical)*). Additionally, a constant, *no_expl_wl*, is employed when no explanation is required. $expl\_wl^k{}_n$ gives the workload of an explanation where k is the time when the scenario occurs and n represents the current time. The formula to calculate this is given in eq. 4.

$$\begin{aligned} min\_wl_n^k &= P\left(not\_critical\right) \cdot no\_expl\_wl \\ &+ \left(1 - P\left(not\_critical\right)\right) \cdot expl\_wl_n^k \end{aligned} \tag{3}$$

$expl\_wl^k{}_n$ is the minimum value of the cost of providing an explanation and the cost of not providing an explanation at time n, given a total scenario duration of k. Taking the minimum reflects the strategic choice of the explanation mechanism, which aims at minimising the expected workload. The cost of providing an explanation depends on the cost of the explanation itself, given as *C(E)*, and the cost which occurs once an explanation is provided (*after_C(E)*). The cost of not providing an explanation varies based on the probability of attention. This value is calculated using the backward Bellman recursion. Herein P(*attn*) is the probability of attention which is obtained by the SEEV model. When attention is being paid, *expl_wl* is the workload cost of pursuing an attention direction by the occupant (*attn_cost*) along with the backward recursion of minimum workload whose horizon is reduced to *k - n*. If there is no attention, then *expl_wl* is the cost of not paying attention (*no_attn_cost*) along with the backward recursion value of the minimum workload.

$$expl\_wl_n^k = min \begin{cases} C(E) + (k-n) \cdot after\_C(E), \\ P(attn)_n \cdot \left(min\_wl_0^{(k-n)} + attn\_cost \right) \\ + (1 - P(attn)_n) \cdot \left(min\_wl_{(n+1)}^k + no\_attn\_cost \right) \end{cases} \tag{4}$$

Based on these rewards, the optimal time to provide an explanation (*t_expl*) and the minimum workload (*min_wl*) for different times until the scenario occurs (*t_max*), is given in the Table 2. Here *t_expl* is the time to provide the explanation from the current instant.

Table 2. Optimal explanation time based on minimum workload.

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

The Table 2 shows the results of optimising explanation timing or horizons *t_max* from *2s* until *10s*. If the horizon is less than *2s*, i.e. if the event occurs within the next second or is already occurring, then the model indicates that no workload reduction can be expected from an early explanation, but this situation changes if the temporal horizon until the event occurs gets larger. Then the exact time of presentation matters: neither presenting as soon as possible nor as late as possible are optimal, but explanation timing is a piecewise affine function of duration of the scenario. In other words, contrary to

intuition it is not best to provide an explanation at the earliest convenience, but there is a defined point in the scenario where it fits best.

When extending scenario durations further, we make the interesting observation that up to scenario duration $t\_max = 15s$, the optimal explanation time is $3s$ before the scenario occurs. But from t_max $= 16s$ onward we see the need to provide two explanations at $2s$ from the start and again at $3s$ before the occurrence of the event. This has already been explored in (Bairy et al., 2023).

Though the backward induction currently is implemented in MATLAB, which does not constitute the most efficient execution platform, we also measured computational runtimes to determine the feasibility of online use of the optimisation procedure for in-situ optimisation of explanation timing in real-time. We found that for smaller values of t_max ($\leq 20s$), the backward induction procedure can be executed online as the computation time stays less than $1s$. But due to the backward recursive function, the computation time exponentially increases. Hence for larger values of t_max, the model needs to be implemented offline or on a more efficient execution platform.

CONCLUSION

This paper introduces the development of a reactive game that utilizes the SEEV model to ascertain the optimal timing for providing explanations. The results presented in the previous section are based on the cost/reward values postulated for the different transitions shown in table 1. These costs currently are just educated guesses serving the purpose of demonstration of the technology, yet lack empirical psychological grounding. We are working together with cognitive psychologists to obtain empirical evidence concerning the actual cost values as well as cross-validation of the optimal timings obtained in table 2, by conducting experiments on real-life subjects.

This research concentrates on determining the optimal timing for providing explanations to a single human present in an AV. Future directions for exploration involve extending the study to scenarios where multiple humans are present, investigating the nuanced dynamics of explanation timing in a group context. Another direction to focus on is the sementic content of an explanation. Rakow and others propose a game-based approach to discerning what content to provide and when (Rakow et al., 2023). The insights gleaned from the present paper could serve as a valuable foundation for the development of such a game, contributing to a more comprehensive understanding of effective explanation strategies for human-AV interaction.

ACKNOWLEDGMENT

The research here has been supported by Universität Oldenburg within the RTG Social Embeddedness of Autonomous Cyber Physical Systems (SEAS) and by Deutsche Forschungsgemeinschaft under grant no. DFG FR 2715/5-1 “Konfliktresolution und kausale Inferenz mittels integrierter sozio-technischer Modellbildung”.

REFERENCES

Baddeley, A., Hitch, G., 1974. Working memory, in: Bower, G. H. (Ed.), Psychology of Learning and Motivation. Elsevier. Volume 8, of Psychology of Learning and Motivation, pp. 47–89. URL: https://doi.org/10.1016/s0079-7421(08)60452-1, doi: 10.1016/s0079-7421(08)60452-1.

Bairy, A., Fränzle, M. (2023). Optimal Explanation Generation using Attention Distribution Model. In: Tareq Ahram and Redha Taiar (eds) Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications. AHFE (2023) International Conference. AHFE Open Access, vol. 70. AHFE International, USA. http://doi.org/10.54941/ahfe1002928

Bairy, A., Hagemann, W., Rakow, A., & Schwammberger, M. (2022), 'Towards Formal Concepts for Explanation Timing and Justifications', 30th IEEE International Requirements Engineering Conference Workshops, RE 2022 – Workshops, Melbourne, Australia, August 15–19, 2022, IEEE. pp. 98–102. URL: https://doi.org/10.1109/REW56159.2022.00025.

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., Robert, L. P., (2019), 'Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload', Transportation Research Part C: Emerging Technologies 104, pp. 428–442, URL: https://www.sciencedirect.com/science/article/pii/S0968090X18313640, doi: https://doi.org/10.1016/j.trc.2019.05.025.

Horrey, W. J., Wickens, C., and Consalus, K. (2006), "Modeling drivers' visual attention allocation while interacting with in-vehicle technologies." Journal of experimental psychology. Applied, 12(2):67–78.

Howard, R. A., 1960. Dynamic Programming and Markov Processes. MIT Press, Cambridge, MA.

Kantowitz, B., 2000. Attention and mental workload. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 44, 3–456. doi: 10.1177/154193120004402121.

Koo, J., Shin, D., Steinert, M., Leifer, L., (2016), 'Understanding driver responses to voice alerts of autonomous car operations', International Journal of Vehicle Design 70, pp. 377. doi: 10.1504/IJVD.2016.076740.

MATLAB, 2022. version 9.13.0 (R2022b). The MathWorks Inc., Natick, Massachusetts.

Rakow, A., Hajnorouzi, M., & Bairy, A. (2023). What to tell when?--Information Provision as a Game. In: Electronic Proceedings in Theoretical Computer Science. FMAS 2023. https://doi.org/10.4204/EPTCS.395.1

Ruijten, P. A. M., Terken, J. M. B., Chandramouli, S., 2018. Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior. Multimodal Technol. Interact. 2, 62. URL: https://doi.org/10.3390/mti2040062, doi: 10.3390/mti2040062.

von Neumann, J., Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press.

Wickens, C., Helleberg, J., Goh, J., Xu, X., & Horrey, W. (2001), 'Pilot Task Management: Testing an Attentional Expected Value Model of Visual Scanning', Savoy, IL, UIUC Institute of Aviation Technical Report.

Wickens, C. (2015), 'Noticing events in the visual workplace: The SEEV and NSEEV models', In The Cambridge Handbook of Applied Perception Research, Cambridge Handbooks in Psychology, pages 749–768. Cambridge University Press.

Wortelen, B., 2014. Das Adaptive-Information-Expectancy-Modell zur Aufmerksamkeitssimulation eines kognitiven Fahrermodells. Ph. D. thesis. Carl von Ossietzky Universität. Oldenburg, Germany.