Human Interaction and Emerging Technologies (IHIET 2023), Vol. 70, 2023, 41–49
https://doi.org/10.54941/ahfe1002928

Optimal Explanation Generation Using Attention Distribution Model

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

With highly automated and Autonomous Vehicles (AVs) being one of the most prominent emerging technologies in the automotive industry, efforts to achieve SAE Level 3+ vehicles have skyrocketed in recent years. As new technologies emerge on a daily basis, these systems are becoming increasingly complex. To help people understand - and also accept - these new technologies, there is a need for explanation. There are three essential dimensions to designing explanations, namely content, frequency, and timing. Our goal is to develop an algorithm that optimises explanation in AVs. Most of the existing research focuses on the content of an explanation, whereas the fine-granularity of the frequency and timing of an explanation is relatively unexplored. Previous studies concerning "when to explain" have tended to make broad distinctions between explaining before, during or after an action is performed. For AVs, studies have shown that passengers prefer to receive an explanation before an autonomous action takes place. However, it seems likely that the acclimatisation that occurs through prolonged exposure to and use of a particular AV will reduce the need for explanation. As comprehension of explanations is workload-intensive, it is necessary to optimise both the frequency, i.e. skipping explanations when they are not helpful to reduce workload, and the precise point in time when an explanation is given, i.e. giving an explanation when it provides the maximum workload reduction. Extra mental workload for passengers can be caused by both giving and omitting an explanation. Every explanation that is presented requires cognitive processing in order to be understood, even if its content is considered to be redundant or if it will not be remembered by the addressee. On the other hand, skipping the explanation can cause the passenger to actively scan the environment for potential cues themselves, if necessary. Such an attention strategy would also impose a significant cognitive load on the passenger. In our work, to predict the mental workload of the passenger, we use the state-of-the-art attention model called SEEV (Salience, Effort, Expectancy, and Value). The SEEV model is dynamically used for forecasting the likelihood of the direction of attention. Our work aims to generate an optimally timed strategy for presenting an explanation. Using the SEEV model we build a probabilistic reactive game, i.e., 1.5-player game or Markov Decision Process, and we use reactive synthesis to generate an optimal reactive strategy for presenting an explanation that minimises workload.

Keywords: Autonomous vehicles, Explanation timing, Reactive game theory, Attention model, Human-Machine-Interaction

INTRODUCTION

As technologies continue to evolve, highly automated and autonomous systems are becoming more complex and less human-understandable. By providing explanations, users can gain a better understanding of how the systems work and become more comfortable interacting with them. This can increase trust and acceptance, which is essential for the successful implementation of these systems in various domains. Game theory (von Neumann & Morgenstern, 1944) being a mathematical model to explain and predict human rational decision making, can also be used to model the policy optimisation for the automated decision making. Recent models from neuropsychology and cognitive psychology suggest that reactive game models, i.e. stateful games, are a better representation of human behavior in decision-making processes. Wickens' Salience, Effort, Expectancy, Value (SEEV) model (Wickens et al., 2001) is a stateful model that helps predict a person's level of attention in a given task or situation.

The main focus of this work is on developing an algorithm to optimize the timing of explanations in autonomous vehicles (AVs) using the SEEV model. To achieve this, the SEEV model is incorporated into a probabilistic reactive game, which is a 1.5-player game or Markov Decision Process. In order to generate an optimal reactive strategy for the rendering of an explanation that minimises the workload, reactive synthesis is used. This paper is a follow-up to preliminary work mentioned in (Fränzle et al., 2023). Our current work concentrates on the fact that all the information required for an explanation might not be available at the start of the scenario, but in fact more information would be obtained over time.

Our paper is organised as follows: in the next section we discuss the related work on explanations in AVs, with a focus on timing of explanations. Then, we give an overview about the SEEV model. Next an example use case is defined followed by the development of our reactive game using the attention obtained from the SEEV model. Finally, an analysis of our experimental results is provided, then a conclusion which includes the limitations of the current model, and the further steps we plan to implement.

RELATED WORK

Shen and others noted that the occupants of an AV do not need explanations in all situations, only in emergencies or near-death (Shen et al., 2020). Much of the existing research focuses on the content and necessity of an explanation, not addressing the potentially critical timing aspects of the explanations.

Research addressing the coarse-granular timing shows that providing an explanation before the occurence of a scenario increases the trust and preference of the system by the occupants of an AV (Koo et al., 2016), (Du et al., 2019), (Ruijten et al., 2018). Koo and others (Koo et al., 2016) examined the effects of explaining a scenario one second before its occurence and their findings show that the participants felt better in control of the situation and more alert.

In their paper, Du and others (Du et al., 2019) examine four different conditions of providing an explanation: before an action, after an action, never providing an explanation and an explanation which was a request for permission. There was an increase in users' trust when explanations were provided before rather than after an action.

For the research of Körber and others (Körber et al., 2018), a situation was set up where the driver was asked to take over and 14s later an explanation was given for asking to take over leading to a better understanding of the situation.

Extending these investigations, we are trying to address the fine-granular timing of explanations given for an imminent event.

SEEV MODEL

SEEV model (*Salience, Effort, Expectancy* and *Value),* originally developed to predict a pilot's attention in a cockpit, helps in quatitatively assessing and predicting the attention level of a human across various areas of interest (Wickens et al., 2001). The four properties of SEEV are: Salience (*S*) – describes how salient new information of a particular type would be to humans if it becomes available. *Effort (*Ef) – refers to the amount of (physical) effort required by the human in order to perceive this new information. *Expectancy (*Ex*)* – refers to how often new information becomes available. It is therefore a dynamic variable that describes the expected remaining time until the arrival of updated information. *Value (*V*)* – is the usefulness of the information item to the human. The formula for calculating the probability of attention P(A) to an item using SEEV is

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

SEEV model is made up of two factors: bottom-up and top-down (Wickens, 2015). Bottom-up factor relates to the physical properties of the environment which affect the attention, i.e. Salience and Effort. Expectancy and value make-up the top down factor. Expectancy changes dynamically based on the time.

EXAMPLE USE CASE

Explanations can consist of multiple pieces of information. The timing of an explanation would be affected based on how much information is available at any given time instant. Let us consider the following example from (Fränzle et al., 2023) to understand this better:

Example scenario 1: *An autonomous vehicle **v** travelling on a road perceives a potential hazard in its path and begins to slow down. The potential danger could be either a cyclist or a deer, but has not been uniquely determined by **v**'s perception components at this stage.*

The Example scenario describes a situation, where the AV has partial information about the particulars of a certain potential hazard at the given time instant. The current information can be provided as an explanation by the AV to the human. This can either lead to a reduction in the cognitive workload on the account of receiving an explanation or an increase in the cognitive workload as the human starts their own attention strategy to fill in the missing information. An alternative option would be to wait until the AV has the complete information as to what the potential hazard is and then provide an explanation. This can again lead to substantial uncertainty as the new information might not be available until it is too late and the human has already started their own attention strategy.

REACTIVE GAME USING SEEV MODEL

As described in the introduction, this note follows up on the submitted article (Fränzle et al., 2023). The reactive game mentioned in that article is modified to take into consideration that explanations are a compilation of information. We build a reactive game graph from the SEEV model and synthesises a strategy to determine the optimal time to provide an explanation, where in contrast to our previous work the explanation can be provided in multiple parts presented at different time instants. The SEEV model describing the probability of attention paid by the human is a function of the evolution of time and the last time of an explanation or an own attentive strategy of the human. This warrants us to develop the reactive game as a Markov Decision Process (MDP), aka 1.5 player game (Howard, 1960). The reactive game consists of a strategic/full player who follows a designated strategy and a random/half player who, based on a given probability distribution, selects an action randomly. In our implementation, the SEEV model constitutes the random player and decides, based on its history-dependent probability, when to pay attention which in turn induces a workload. The explanation mechanism represents the strategic player and strategically decides whether the explanation should better be provided once the whole information is present or before in order to minimise the expected cognitive workload on the human as Bairy and others suggest in their paper (Bairy et al., 2022). It should be noted that the presentation of an explanation also induces workload on the person associated with the reception and interpretation of the explanation, although generally at a lower level of workload than the pursuit of an active attention strategy.

In our game, we aim to find the optimal time for explanation(s) in Example scenario 1. Just like the example scenario presented in (Fränzle et al., 2023), there is only a fixed area of interest. Thus the effort factor remains constant. We can also approximate salience as being constant throughout the scenario, given the short time span of the scenario. Thus both *Salience* and *Effort* and consequently also their difference can be replaced by a constant in equation (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 reactive game with the SEEV model, as a dynamic factor impacting workload within a 1.5 player game, to compute the reactive presentation strategy, has been implemented in MATLAB (MATLAB, 2022).

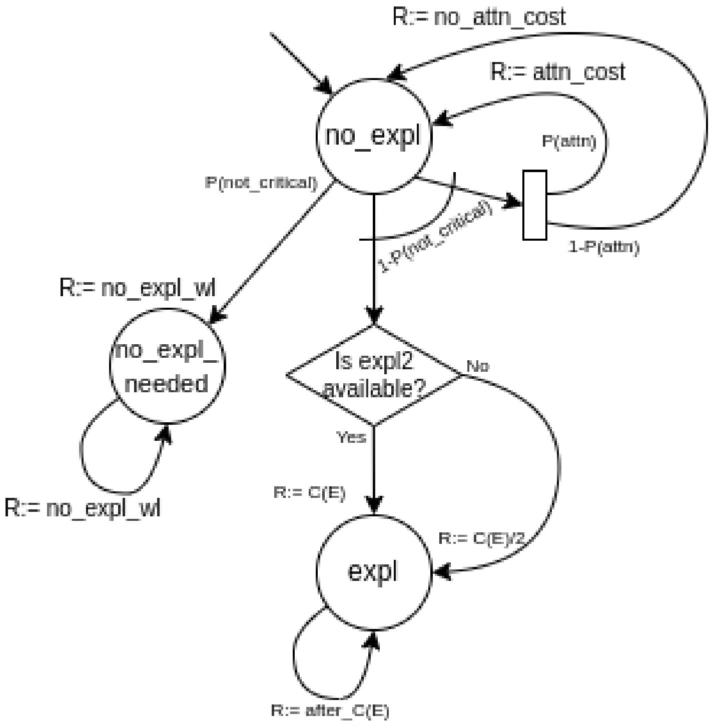

Figure 1: State diagram of the strategic player with two explanations.

The explanation is assumed to be consisting of two parts of information. The reactive SEEV game starts *n* time seconds before the scenario occurs, and ends when the scenario finalises. The first part of the information is considered to be available *n* seconds before the scenario occurs. The second part of the information in our example scenario becomes available *n*/2 seconds before the scenario occurs.

At each time step, which is set to one second granularity, the strategic player has an option to execute an action that puts it in one of the following three states: *no_expl* which represents not providing an explanation; *expl* which provides an explanation; and *no_expl_needed* which indicates that an explanation might not be necessary at certain times (Fränzle et al., 2023). Each of the strategic player's {state, action} pairs is associated with certain costs/rewards. The attention level of the occupant is measured using equation (2) based on the strategic player's chosen action and its associated cost. A state diagram for the strategic player's different states and their transitions is shown in Figure 1.

Figure 2 shows the build up of expectancy of the random player over time. It builds up with a constant value of *exp*. Since there is only one area of interest in our model, as we are implementing it for only the example scenario 1, the value aspect of the SEEV model can also be considered a constant. There are costs/rewards for various actions undertaken by the strategic player. The probability of attention (*P(attn)*) and the probability of no critical

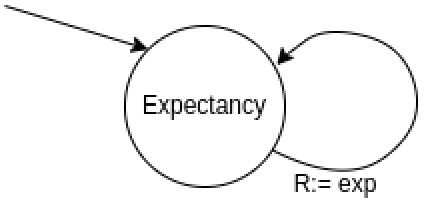

Figure 2: Expectancy of the random player (Fränzle et al., 2023).

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 (after expl2 is available)	0.1
expl	expl	1 (before expl2 is available)	0.05
no_expl_needed	no_expl_needed	1	0.0

scenario (*P(not_critical)*) help in determining the reward values for the strategic player. *P(attn)* is calculated using equation (2). The details about the reward structure is discussed more in the next section.

EXPERIMENTAL RESULTS

Our goals of this reactive SEEV game were to identify, given that the AV determines information contributing to the explanation asynchronously in two parts, would the explanation need to be provided in two parts and what then is the optimal time to provide the explanation or explanation part in such a manner that the cognitive workload of the human is minimised. A cost structure detailing cognitive workload costs was assigned to the various state transitions to calculate this. The cost structure is shown in Table 1.

The current model assumes a fixed time until the scenario occurs, i.e. a finite horizon. Hence the minimum workload induced by the attention strategy under an optimal strategy for explanation timing can be determined using backward Bellman induction. Equation (3) gives the minimum workload (*min_wl*) (Fränzle et al., 2023). Probability of no critical scenario (*P(not_critical)*) ensures that a situation where the user might have already evaluted the surroundings or the critical situation was resolved by the temporal evolution of the situation is also accounted for. A constant workload value associated with *P(not_critical)* is given by *no_expl_wl*. In scenarios where an explanation might be required, workload ($expl\_wl^k{}_n$) is calculated using backward induction with k as the time when the scenario occurs and n being the current time. Equation 4 shows how this is computed.

$$min\_wl_n^k = P\left(not\_critical\right) \cdot no\_expl\_wl + \left(1 - P\left(not\_critical\right)\right) \cdot expl\_wl_n^k \qquad (3)$$

If the total scenario time is k, $expl\_wl^{k}{}_{n}$ is the minimum between the costs incurred when explaining now and the costs incurred when not explaining, reflecting the strategic decision that minimises the expected workload. Waiting for an explanation comes with a cost (waiting_cost) which depends on the probability of attention (*P(attn)*) obtained from SEEV model in equation (2). This cost is again calculated using backward Bellman recursion. When the human is paying attention, waiting_cost becomes the cost of pursuing an attention direction by the occupant (*attn_cost*) combined with a minimum workload with a reduced horizon (k-n) (Fränzle et al., 2023). When the human is not paying attention, waiting_cost is the cost of not paying attention (*no_attn_cost*) combined with the minimum workload obtained by backward recursion (Fränzle et al., 2023).

$$expl\_wl_{n}^{k} = min \begin{cases} expl\_cost, \\ P\left(attn\right)_{n} \cdot \left(\min\_wl_{0}^{(k-n)} + attn\_cost\right) \\ + \left(1 - P\left(attn\right)_{n}\right) \cdot \left(\min\_wl_{(n+1)}^{k} + no\_attn\_cost\right) \end{cases} \tag{4}$$

Providing an explanation comes with a cost (*expl_cost*) shown in equation (5). *expl_cost* is the sum of the costs of receiving an explanation (*C(E)*) and a constant cost which occurs after an explanation is received (*after_C(E)*) (Fränzle et al., 2023). Since in our model there are two explanations available, explanation with partial information at the start and complete information explanation available later, the cost/reward of explanation depends on which explanation is being provided. If the first (partial) explanation is provided, then only half the reward *C(E)* is awarded. Table 2 gives an overview of the minimum workload(s) (*min_wl*) and optimal time(s) (*t_expl1/t_expl2*) to provide explanation(s) for different times until the scenario occurs (*t_max*) based on the above mentioned reward structure.

$$expl\_cost = \begin{cases} C\left(E\right) + \left(k-n\right) \cdot after\_C\left(E\right) & expl2\ is\ available \\ 0.5 \cdot C\left(E\right) + \left(k-n\right) \cdot after\_C\left(E\right) & otherwise \end{cases} \tag{5}$$

Table 2 shows the results of optimising explanation timing for horizons t_max from 2s until 15s. For events starting within a second or already occuring the model indicates that no workload reduction can be expected from an early explanation. However, this situation changes as the time to the event increases. Then the exact timing of the explanation is important: neither the earliest nor the latest is optimal, but the timing of the explanation is a piecewise affine function of the duration of the scenario. That is, contrary to what might seem intuitive, it is not best to explain as soon as possible, but there is a defined point during the scenario when it is best to do so. Up until 9s horizon, even though a part of the explanation is available earlier, an explanation is required closer to the situation, namely 3s before the scenario occurs, i.e. when all the information is available. But this changes from 10s duration onwards. A partial explanation is helpful when presented at 2s and then again another completing explanation is expected 3s before the scenario occurs.

Table 2 also shows the computation runtimes for the scenario with different *t_max*. Due to the backward induction the computation runtime exponentially increases in the length of the horizon. But even with MATLAB,

Table 2. Optimal explanation times for 2 explanations based on minimum workload.

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

which is far from being the most efficient execution platform, the computational time is less than 1s for *t_max* values until 15s and hence the algorithm can be executed online. Higher *t_max* values may require it to be executed offline or on a more efficient execution platform.

CONCLUSION

In this paper, we developed a reactive game using the SEEV model to determine the optimal time to provide explanation(s), if the explanation comes in chunks: initial explanation with only partial information being possible at the start, and a second explanation with complete information becoming available later.

The results obtained in Table 2 are based on costs which are educated guesses designed to demonstrate the technology, but are not based on empirical psychological research. Such an empirical evaluation and parameterisatoin remains to be done.

This research only focuses on the timing of the explanation(s). The next step would be to include the sementic content of the parts of an explanation and to observe how the optimal timing of an explanation might change when certain combinations are more comprehensible than others due to causal dependencies.

ACKNOWLEDGMENT

The research reported here has been supported by Deutsche Forschungsgemeinschaft under grant no. DFG FR 2715/5-1 “Konfliktresolution und kausale Inferenz mittels integrierter sozio-technischer Modellbildung”, by the State of Lower Saxony within the Zukunftslabor Mobilität.

REFERENCES

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

Fränzle, M., Bairy, A., & Hajnorouzi, M. (2023), 'Computational Cognitive Models Meet Reactive Game Theory and Reactive Synthesis: Cognitively Informed Automated Synthesis of Behavioural Strategies Across the Human-Machine Boundary', International Journal of Human-Computer Studies (Submitted)

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

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

Körber, M., Prasch, L., Bengler, K., (2018), 'Why do I have to drive now? Post hoc explanations of takeover requests', Human Factors 60, pp. 305–323, doi:10.1177/0018720817747730.864

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

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

Shen, Y., Jiang, S., Chen, Y., Yang, E.J., Jin, X., Fan, Y., & Campbell, K.D. (2020), 'To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles', ArXiv, abs/2006.11684

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: {T}esting 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 University Press. Cambridge Handbooks in Psychology, pp. 749–768. doi:10.1017/CBO9780511973017.046.