AI for Explaining Decisions in Multi-Agent Environments*

Sarit Kraus,[1] Amos Azaria,[2] Jelena Fiosina,[3] Maike Greve,[4] Noam Hazon,[2]
Lutz Kolbe,[4] Tim-Benjamin Lembcke,[4] Jörg P. Müller,[3] Sören Schleibaum,[3] Mark Vollrath[5]

[1]Department of Computer Science, Bar-Ilan University, Israel (email: sarit@cs.biu.ac.il)
[2]Department of Computer Science, Ariel University, Israel
[3]Department of Informatics, TU Clausthal, Germany
[4]Chair of Information Management, Georg-August-Universitat Göttingen, Germany
[5]Chair of Engineering and Traffic Psychology, TU Braunschweig, Germany

Abstract

Explanation is necessary for humans to understand and accept decisions made by an AI system when the system's goal is known. It is even more important when the AI system makes decisions in multi-agent environments where the human does not know the systems' goals since they may depend on other agents' preferences. In such situations, explanations should aim to increase user satisfaction, taking into account the system's decision, the user's and the other agents' preferences, the environment settings and properties such as fairness, envy and privacy. Generating explanations that will increase user satisfaction is very challenging; to this end, we propose a new research direction: Explainable decisions in Multi-Agent Environments (xMASE). We then review the state of the art and discuss research directions towards efficient methodologies and algorithms for generating explanations that will increase users' satisfaction from AI system's decisions in multi-agent environments.

Introduction

Many AI systems need to make decisions in multi-agent environments where the agents, including people and robots, have possibly conflicting preferences. The system should balance between these preferences when making decisions regarding all agents. Such systems include, e.g., a scheduling algorithm assigning teachers to classes, or a ridesharing application proposing joint rides to people. In such situations, the global decisions made by the system may not adhere to all people's preferences: a decision may make some people unhappy. Providing explanations about the system's decision may increase people's satisfaction (Bradley and Sparks 2009), and maintain acceptability of the AI system (2018). The EU General Data Protection Regulation introduces a right of explanation (Goodman and Flaxman 2017) for citizens to obtain "meaningful information of the logic involved" for automated decisions.

Explainable AI (XAI) has recently been studied extensively (Core et al. 2006; Carvalho, Pereira, and Cardoso 2019; Rosenfeld and Richardson 2019), mainly focusing on finding ways of explaining to a user a decision made by an AI system, e.g., a classification choice made by a neural network. That is, explanations are usually given for black-box algorithms, which aim at maximizing a well-agreed upon function (e.g., maximizing accuracy, minimizing loss). Previous work has mainly focused on increasing users' trust in the black-box AI system. However, we believe that providing the users with explanations is even more important in multi-agent environments, when even the maximization function is not clear to the (human) agents. In such situations, explanations should aim to increase user satisfaction, taking into account properties such as fairness, envy and privacy. Hence, we propose a new research direction of Explainable decisions in Multi-Agent Environments (xMASE).

For example, in the ridesharing domain, an AI system may suggest to a customer Bob to share a taxi with Alice in a ride that will take 30 minutes. The taxi will first drop Alice off and it will cost Bob $25. Bob may be upset that the taxi drops Alice off first. An explanation could be that Alice's destination is on Bob's route and will only add 5 minutes to his trip. The system can also say that Alice will pay $30, or that dropping Bob off first will add 15 minutes to Alice's trip, or that Alice teaches at 8 am and will be late if Bob is dropped first. Bob can also be told that sharing a taxi will save him $10.

Generating explanations in multi-agent environments is even more challenging than providing explanations in other settings, e.g., for classification results produced by deep learning algorithms. In addition to identifying the technical reasons that led to the decision, there is a need to convey the preferences of the agents that were involved. It is necessary to decide what to reveal from other agents' preferences in order to increase the user's satisfaction, taking the privacy of other agents into account, and how these preferences led to the final decision. It should also refer to issues such as fairness.

The influence of the explanation on user satisfaction changes from one user to the next; therefore, personalized explanations are beneficial (Lakkaraju et al. 2019; Bradley and Sparks 2009). Given that the task is very challenging, we propose to use AI tools, and in particular machine learning to generate the personalized explanations that will maximize user satisfaction, while aiming to consider system-level societal welfare aspects. In order to use machine-learning methods for generating explanations in our context, it will be necessary to collect data about human satisfaction from decision-making when various types of explanations are given in different contexts. Furthermore, the evaluation of the proposed methods must also involve humans participating in multi-agent environments.

Most algorithms that provide explanations on AI systems take an engineering approach, which does not involve running experiments with people. Papenmeier et al. (2019) showed that the presence of such explanations did not increase the subjects' self-reported trust. Miller (2018) argues that explanations have been studied extensively in psychology and their findings should be used when designing explanations for AI systems. One of his main points is that an explanation should be sensitive to context. We fully agree that for both XAI and xMASE context should be taken into consideration when generating explanations, but in multi-agent systems (MASs) context includes other agents' preferences, and fairness of the decision as an important factor.

For example, consider a scheduling algorithm that assigns teachers to classes. Suppose Bob and Alice each teach in a specific classroom on a specific day; each needs to teach for 4 hours (but they cannot teach in parallel). Bob prefers to teach between 10am and 3pm, but can also teach from 9am-10am and 3pm-4pm. He cannot teach between 8am and 9am (a strong constraint) or after 4pm. Alice prefers to teach between 10am-2pm, has a hard constraint not allowing her to teach after 2pm, but can teach in the mornings, 8am-10am. Suppose the algorithm assigned Bob to 12pm-4pm and Alice to 8am-12pm (see Figure 1). Which explanations should the system provide for Alice whose assignment violates her soft constraint but not her hard constraints? Which explanations should the system give a teacher for an assignment violating his or her hard constraints? Suppose Alice and Bob are friends. Should the system tell Bob that he teaches between 3pm and 4pm (violating his soft constraint) because Alice has a strong constraint at this time? What if Alice and Bob are new to the school and do not know one another; should such an explanation be provided? Alternatively, could a merely graphical description of the relevant information (as in Figure 1) suffice as an explanation?

There are many challenging research questions to be studied in order to provide users in MASs with explanations that will increase their satisfaction:

Algorithms for Explanation Generation: The develop-

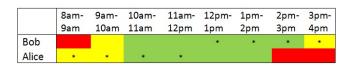|	8am-9am	9am-10am	10am-11am	11am-12pm	12pm-1pm	1pm-2pm	2pm-3pm	3pm-4pm
Bob					*	*	*	*
Alice	*	*	*	*				

Figure 1: Scheduling example assigning teachers to classes, depicting Alice and Bob's preferences (green=preferred, yellow=feasible, red=impossible) and assignments (*).

ment of efficient algorithms that will generate the explanations, preferably in real time, is very challenging. First, there is a need to be able to form good explanations. Then, there is a need to decide which explanation to present (if at all) and when and how to present it.

User modeling for increasing satisfaction: Most important is to be able to model the users. It is, of course, necessary to identify the users' preferences in order to make good decisions. However, for generating good explanations it is important (beyond obtaining the users' preferences) to also model their attitudes toward different explanations. That is, to be able to predict how an explanation will influence users' satisfaction.

Interactive explanations: To address the difficulties in choosing the specific explanation to a specific user, we propose to also study interactive explanations, which are provided to the user through a dialogue between the AI system and the user (Miller 2018). By asking questions and expressing his or her concerns, the user can direct the system towards generating good explanations. However, conducting meaningful dialogues with a user for xMASE adds difficulties to the process. It has some similarities to automated systems that argue with people, a research area that still has many open questions (Rosenfeld and Kraus 2016a; Rosenfeld and Kraus 2016b).

Understanding System Decision-making: The decision of an xMASE system depends on many parameters associated with several agents, making the problem even more challenging than in XAI. Only some of the technical reasons that led to the decision are relevant to a given user; as the number of agents increases, the non-relevant information increases, too, and it is hard to identify the relevant parts. Other explanations that could increase user satisfaction concern the environment. If the AI system made a decision based on some knowledge about the environment, but this knowledge is not available to the user, presenting it to the user could be useful. E.g., if Bob believes that there are many taxis available at 8am, but the system proposed him to share a taxi with Alice since it knows that there are very few taxis available, giving this information to Bob might increase his satisfaction.

Long-term relationships: When the AI system inter-

acts repeatedly with the same users, interesting research questions may arise. The learning phase of the preferences and satisfaction models can be personalized, but more importantly the explanation generated should take long-term satisfaction into consideration.

Ethics and Privacy: These issues must be considered when presenting explanations in multi-agent systems. As in other situations, one needs to regard issues such as the truthfulness of the explanations and the ethical aspect of concealing some of the information. Moreover, privacy is a major concern in xMASE, since there is also a need to consider the revelation of information and preferences of one agent to another when providing an explanation.

Open source code and public datasets: Open source environments that will allow researchers to develop and evaluate their explanation algorithms can enhance the research in xMASE. It is also very useful to start collecting labeled datasets for xMASE.

We will discuss three of these directions in more detail after surveying the related work to xMASE.

State-of-the-art

In recent years there has been intense research on design techniques for making AI methods explainable, interpretable, and transparent for developers and users (Carvalho, Pereira, and Cardoso 2019). The basic idea of XAI methods is to try to explain black-box model behaviour (while xMASE is needed even in settings where traditional white-box optimization is used). Methods for XAI have been developed including locally interpretable model-agnostic explanations for Bayesian predictive models (Peltola 2018) and for convolutional neural networks (Mishra, Sturm, and Dixon 2017), visualization techniques (Grad-CAM) for CNNs (Selvaraju et al. 2017), or black box explanations through transparent approximations (Lakkaraju et al. 2017). Hybrid models use explicit symbolic representations in conjunction with black-box techniques (Hu et al. 2016; Choi, Wang, and Darwiche 2019). XAI was also expanded beyond the classical domains. For example, Fox et al. (2017) introduces explainable planning systems. Ludwig et al. (2018) and Čyras et al. (2019) investigated the explainability of a scheduling system over tasks. We note that both planning and scheduling of tasks to machines are not multi-agent environments as we define them, since they actually consist of a single agent, and therefore do not belong to xMASE. The explainability of deep reinforcement learning was also investigated (Lee 2019). All of these techniques were developed for XAI, and they can serve as input for xMASE that needs to choose the suitable explanation for any given scenario and to each agent. Moreover, in many XAI approaches the evaluation is complicated due to missing standards (Pedreschi, Giannotti, and others 2019), the social nature of explanations is ignored (Miller 2018), and the explanations are not eval-

uated with humans. All of these evaluation criteria are essential in xMASE.

Indeed, there are some works which provide evaluations of a XAI approach with humans. Lakkaraju et al. (2019) propose a new form of explanations designed to help end users (e.g., decision-makers such as judges, doctors) to gain a deeper understanding of the models' behaviour. Doshi-Velez and Kim (2018) claim that researchers evaluating explainability should differentiate between evaluation of explanations with humans, experts in the evaluated fields, and a formal evaluation without human subjects. Wolf et al. (2019) point out requirements for explanations and state that these should be influenced by the users, the applications and the deployment context. They claim that three different types of explanations are needed for integrating XAI approaches into real world applications. Firstly, an application should explain its behavior; Secondly, the impact of the interaction of users with the applications should be explained, and lastly, a user seeks explanations describing how the applications' output integrates in the overall process. These ideas are also some of the basic ingredients of a successful xMASE.

One unique aspect of xMASE is that the explanations should be chosen such that they will increase the user's satisfaction. Previous works have shown that explanations in general have an impact on user satisfaction/acceptance. Herlocker et. al. (2000) showed that providing explanations for automated collaborative filtering (ACF) recommendations can improve the acceptance of ACF systems. More relevant to us is (Kleinerman, Rosenfeld, and Kraus 2018) showing that explanations are beneficial also in reciprocal recommendation systems (e.g., dating) where the preferences of the other agent are taken into account when making recommendations, and possibly when generating explanations. In such settings reciprocal explanations are preferable. Putnam and Conati (2019) conducted a user study on benefits from explanations in intelligent tutoring systems. Their results indicate a positive sentiment towards wanting explanations, but do not suggest any automatic system for their generation. Levinger et al. (2018) study maximizing human satisfaction in multi-agent systems, but do not use any form of explanation. They present an optimization algorithm that maximizes the overall human satisfaction according to the learned model. They also show that, when aiming at maximizing human satisfaction, learning accurate models of human satisfaction is more important than improving the optimization algorithm.

Research directions for xMASE

The development of AI-based tools that provide the right explanations to the right users at the right time to increase user satisfaction in MASs is very challenging. We now discuss three of the above research directions in more detail.

Generation of explanations to increase satisfaction

The development of efficient algorithms that will generate the explanations, preferably in real time, is very important. We propose a two stage procedure: first, a set of possible explanations will be created and then the one that best suits the specific user at the specific settings will be selected. Both stages can be done using machine learning or any other decision-making procedures based on real user input.

If the AI decision is made using neural networks (e.g., (Rosemarin, Rosenfeld, and Kraus 2019; Li, Qin, and others 2019)) then XAI methods can be used to identify important features that led to the decision (Shrikumar, Greenside, and Kundaje 2017; Bach, Binder, and others 2015). These methods should be adapted to xMASE-related problems (Lee 2019; Selvaraju et al. 2017). Then, there is a need to identify which of these features are relevant to a specific user. Given these features, the preferences of other agents that are relevant should be identified and any relevant statements that touch upon important concepts such as fairness should be generated. Using these features, preferences and concepts, several explanations could be generated using subsets of them. Next, the explanations with the relevant preferences and environment settings could be entered into a network that will estimate the influence level of each explanation on the user's satisfaction. Finally, the chosen subset should be transferred into a textual message and sent to the user. Personalization could also be used in this final step.

If the AI decision-making is not done using machine learning methods, but rather, e.g., using inferences, and data is not available, it will be interesting to study how methods developed for explaining inferences (e.g., (Pino-Pérez and Uzcátegui 2003)) could be used for xMASE.

If the AI decision-maker is a scheduling tool, it can provide a set of constraints that lead to the proposed schedule (Ludwig, Kalton, and Stottler 2018). In xMASE there will be a need to identify the relevant constraints and to generalize statements related to other agents' preferences, and the general system constraints that are driven by other concepts such as fairness. Then again, we can use user satisfaction models (e.g., represented by a neural network) to choose the best constraints and generalized statements to be presented.

Complementing the algorithms and methodologies described so far, an interesting research direction is the automated generation of graphical explanations for xMASE, enabling compact summarization of large amounts of information.

User modeling for increasing satisfaction

There are many methods for user modeling and preference elicitation (Rosenfeld and Kraus 2018; Anselmi et al. 2018), but we have not found a study on preference elicitation with respect to explanations and their role in improving satisfaction. One of the main challenges is that user satisfaction from an explanation of a given decision strongly depends on the actual decision, the other agents, the environment and the user's beliefs. For example, it is obvious that, in our ridesharing example, an explanation to Alice of the form "This shared ride will save you $10" is a "good" explanation, but of course if in the specific ride Alice will save time but not money, the explanation is useless. Similarly, telling Bob that "There is no private taxi available in your area now" may be a good explanation for a ridesharing suggestion, given that it is true (assuming we provide only true explanations) and Bob does not know it. Thus collecting data on the influence of an explanation on the user's satisfaction must be done in the context of the specific decision it explains and the environment setting. This makes data collection very challenging.

Collection of data can be done either using fictitious decisions, their explanations and the MAS environment setting or, much harder to accomplish, in actual settings or at least in simulations. The users can express their preferences on how much they like the explanations. We can use this data to build a generalized model of users' preferences toward explanations. However, this model will not provide us with the explanations that increase the user's satisfaction. Here we will need to let the user express his or her level of satisfaction from a given decision with different variants of explanations and without explanations, and try to build a model that measures the users' satisfaction from the decision.

One of the challenges is to identify features of an explanation. We may consider using the explanation as text, but we believe that there is a need to find additional features that have to do with the relationship of the explanation with the user's preferences and the environment.

When the AI system interacts repeatedly with the same users, interesting research questions may arise. The learning phase of the preferences and satisfaction models can be personalized, but more importantly the explanation generated should take long-term satisfaction into consideration. Furthermore, we propose to consider, when interacting with the user, using reinforcement learning to improve the user's model of overtime in a guided way.

Interactive explanations

Human verbal explanations are essentially interactive (Cawsey 1993). Recently, there have been a few attempts to consider models for interactive explanations to XAI (Madumal et al. 2018) and to value-based agents (Liao, Anderson, and Anderson 2018), but no system was developed. Interactive explanations can be viewed as argumentation dialogues. It was shown to be beneficial to model the interaction as POMDP where the uncertainty is about the user's beliefs (Rosenfeld and Kraus 2016b). Using this approach for xMASE there is a need to continuously estimate the user's beliefs and sentiment toward the AI system's decision, and to pre-

dict how a given explanation statement will modify the user's beliefs and influence its attitude toward the decision.

Other techniques for general dialogue systems or for interactive learning dialogues could be considered (Chen and others 2017). The open questions are how to (1) use the features that led the AI system to make a decision to form possible responds, and (2) devise user models of preferences and satisfaction to generate the right response in the dialogue. Regardless of the techniques, any developed method should be evaluated via human experiments, which are still often missing in XAI.

and acceptability of socio-technical applications such as ridesharing where people's preferences may conflict, but XAI's recent progress, but many open questions should the environment and the goal of increasing user's satis

References

[Anselmi et al. 2018] Anselmi, P.; Fabbris, L.; Martini, M. C.; and Robusto, E. 2018. Comparison of four common data collection techniques to elicit preferences. Quality & Quantity 52(3):1227–1239.

[Bach, Binder, and others 2015] Bach, S.; Binder, A.; et al. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7).

[Bradley and Sparks 2009] Bradley, G. L., and Sparks, B. A. 2009. Dealing with service failures: The use of explanations. Journal of Travel & Tourism Marketing 26(2):129–143.

[Carvalho, Pereira, and Cardoso 2019] Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. ML interpretability: A survey on methods and metrics. Electronics 8(8):832.

[Cawsey 1993] Cawsey, A. 1993. Planning interactive explanations. Intern. J. of Man-Machine Studies 38(2):169–199.

[Chen and others 2017] Chen, H., et al. 2017. A survey on dialogue systems: Recent advances and new frontiers. ACM SIGKDD Explorations Newsletter 19(2):25–35.

[Choi, Wang, and Darwiche 2019] Choi, A.; Wang, R.; and Darwiche, A. 2019. On the relative expressiveness of bayesian and neural networks. International Journal of Approximate Reasoning.

[Core et al. 2006] Core, M. G.; Lane, H. C.; Van Lent, M.; Gomboc, D.; Solomon, S.; and Rosenberg, M. 2006. Building explainable artificial intelligence systems. In AAAI-06.

[Čyras et al. 2019] Čyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019. Argumentation for explainable scheduling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2752–2759.

[Doshi-Velez and Kim 2018] Doshi-Velez, F., and Kim, B. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning. In Explainable and Interpretable Models in Computer Vision and ML. 3–17.

[Fox, Long, and Magazzeni 2017] Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable planning.

[Goodman and Flaxman 2017] Goodman, B., and Flaxman, S. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine 38(3):arXiv:1606.08813.

[Herlocker, Konstan, and Riedl 2000] Herlocker, J. L.; Konstan, J. A.; and Riedl, J. 2000. Explaining collaborative filtering recommendations. In CSCW, 241–250. ACM.

[Hu et al. 2016] Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. 2016. Harnessing deep neural networks with logic rules. In ACL, volume 1, 2410–2420.

[Kleinerman, Rosenfeld, and Kraus 2018] Kleinerman, A.; Rosenfeld, A.; and Kraus, S. 2018. Providing explanations for recommendations in reciprocal environments. In ACM Recommender Systems, 22–30. ACM.

[Lakkaraju et al. 2017] Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2017. Interpretable & explorable approximations of black box models. In FAT ML (KDD).

[Lakkaraju et al. 2019] Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2019. Faithful and customizable explanations of black box models. In AIES.

[Lee 2019] Lee, J. H. 2019. Complementary reinforcement learning towards explainable agents. CoRR abs/1901.00188.

[Levinger, Azaria, and Hazon 2018] Levinger, C.; Azaria, A.; and Hazon, N. 2018. Human satisfaction as the ultimate goal in ridesharing. arXiv preprint arXiv:1807.00376.

[Li, Qin, and others 2019] Li, M.; Qin, Z.; et al. 2019. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In WWW, 983–994.

[Liao, Anderson, and Anderson 2018] Liao, B.; Anderson, M.; and Anderson, S. L. 2018. Representation, justification and explanation in a value driven agent: An argumentation-based approach. arXiv preprint arXiv:1812.05362.

[Ludwig, Kalton, and Stottler 2018] Ludwig, J.; Kalton, A.; and Stottler, R. 2018. Explaining complex scheduling decisions. In IUI Workshops.

[Madumal et al. 2018] Madumal, P.; Miller, T.; Vetere, F.; and Sonenberg, L. 2018. Towards a grounded dialog model for explainable artificial intelligence. arXiv preprint.

[Miller 2018] Miller, T. 2018. Explanation in artificial intelligence: Insights from the social sciences. AIJ 267:1–38.

[Mishra, Sturm, and Dixon 2017] Mishra, S.; Sturm, B. L.; and Dixon, S. 2017. Local interpretable model-agnostic explanations for music content analysis. In ISMIR.

[Papenmeier, Englebienne, and Seifert 2019] Papenmeier, A.; Englebienne, G.; and Seifert, C. 2019. How model accuracy and explanation fidelity influence user trust. In 3rd Workshop on Explainable AI (X-AI), 94–100.

[Pedreschi, Giannotti, and others 2019] Pedreschi, D.; Giannotti, F.; et al. 2019. Meaningful explanations of Black Box AI decision systems. In AAAI, volume 33, 9780–9784.

[Peltola 2018] Peltola, T. 2018. Local interpretable model-agnostic explanations of bayesian predictive models via kullback-leibler projections. In X-AI@IJCAI.

[Pino-Pérez and Uzcátegui 2003] Pino-Pérez, R., and Uzcátegui, C. 2003. Preferences and explanations. Artificial Intelligence 149(1):1–30.

[Putnam and Conati 2019] Putnam, V., and Conati, C. 2019. Exploring the need for explainable artificial intelligence (xai) in intelligent tutoring systems (its). In IUI Workshops.

[Rosemarin, Rosenfeld, and Kraus 2019] Rosemarin, H.; Rosenfeld, A.; and Kraus, S. 2019. Emergency department online patient-caregiver scheduling. In AAAI-19.

[Rosenfeld and Kraus 2016a] Rosenfeld, A., and Kraus, S. 2016a. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. ACM Trans. on Interactive Intelligent Systems (TiiS) 6(4):30.

[Rosenfeld and Kraus 2016b] Rosenfeld, A., and Kraus, S. 2016b. Strategical argumentative agent for human persuasion. In ECAI, 320–328.

[Rosenfeld and Kraus 2018] Rosenfeld, A., and Kraus, S. 2018. Predicting human decision-making: From prediction to action. Synthesis Lectures on Artificial Intelligence and Machine Learning 12(1):1–150.

[Rosenfeld and Richardson 2019] Rosenfeld, A., and Richardson, A. 2019. Explainability in human–agent systems. JAAMAS.

[Selvaraju et al. 2017] Selvaraju, R. R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization. In ICCV-17.

[Shrikumar, Greenside, and Kundaje 2017] Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In ICML-17, 3145–3153.

[Wolf, Blomberg, and Jose 2019] Wolf, C. T.; Blomberg, J. L.; and Jose, S. 2019. Explainability in context: Lessons from an intelligent system in the it services domain.