A Communicative Action Framework for Discourse Strategies for AI-based Systems: The MetTrains Application Case

Research in Progress

Shirley Gregor
Australian National University
shirley.gregor@anu.edu.au

Alexander Maedche
Karlsruhe Institute of Technology
alexander.maedche@kit.edu

Stefan Morana
Karlsruhe Institute of Technology
stefan.morana@kit.edu

ABSTRACT

Increasing attention is being paid to the challenges of how artificial intelligence (AI)-based systems offer explanations to users. Explanation capabilities developed for older logic-based systems still have relevance, but new thinking is needed in designing explanations and other discourse strategies for new forms of AI that include machine learning. In this work-in-progress paper we show how a communicative action design framework can be used to design an AI-based system's interface to achieve desired goals. The applicability of the framework is demonstrated with an interface for an intelligent video surveillance system for reducing railway suicide. The communicative action framework is an important step in theory development for human-computer interaction with AI as used in the information systems domain.

Keywords

Artificial intelligence, discourse strategy, explanations, human-computer interaction, communicative action

INTRODUCTION

The deployment of artificial intelligence (AI)-based systems in organizations and for personal use is increasing rapidly and contributing significantly to the revolutionary transformations occurring in our societies. Thus, interactions between humans and AI-based systems become increasingly important (Brynjolfsson & McAfee, 2016). However, there are ongoing issues with AI-human interaction, allowing humans to understand and trust the actions taken by AI-based systems and engage in collaborative decision making. Providing explanations is one form of support for users' trust and understanding of intelligent systems (Gregor & Benbasat, 1999). However, providing explanations to justify an AI's recommendations is difficult with machine learning (ML) systems now in common use, leading to a potential lack of trust by users in such "black boxes" (Knight, 2017; Guidotti *et al.*, 2019). There is an accompanying increase in interest in what is now termed "explainable AI" (Mittelstadt *et al.*, 2019; Gunning, 2018). However, it is not clear that lessons learned in the past with explanations in older forms of AI, such as expert systems, are being heeded as much as they could be. On the other hand, full transparency from explanations may not be required when the main goal is efficiency and effectiveness and may even be harmful (e.g. see Weller, 2017). Further, going beyond explanations, some systems may utilize influencing techniques to guide users in certain ways and some may invoke affective responses to aid the communication process.

Against this background we propose that further theory development is needed for the design of the discourse strategies that are evidenced in the interaction of AI with human users. Specifically, we propose theory building that draws on Habermas' theory of communicative action (Habermas, 1984) as well as prior theoretical and empirical work. In this work-in-progress paper we demonstrate how our new communicative action (CA) framework can be used to improve the design of an AI-based system's interface, with an exemplar case of an intelligent video surveillance system aimed at reducing railway suicide in a metropolitan railway network (MetTrains[1]).

The paper proceeds by giving an overview of the literature on explanations and other discourse strategies for AI-based systems. This is followed by an outline of the new CA design framework. The context of the application case is then described, including both the railway suicide case study background and the capabilities of intelligent surveillance systems. The application of the CA framework follows, along with our conclusions.

[1] A pseudonym.

CONCEPTUAL BACKGROUND – AI-BASED SYSTEMS AND DISCOURSE STRATEGIES

We use the concept of AI-based systems in a broad sense to include systems with a range of abilities, congruent with prior concepts such as intelligent system (Gönül *et al.*, 2006; Gregor & Benbasat, 1999). We follow Gregor & Benbasat (1999) in defining intelligent systems as "information systems with an "intelligent" or "knowledge component". In the remainder of the paper, we refer to AI-based systems as subsuming all types of systems, such as expert systems, decision support systems (DSS), recommender agents, conversational agents, and business intelligence and analytics systems independent of the underlying technology.

Users and AI-based systems can work together to engage in intelligent activities. To do so, they require specific functionalities that allow for the communication between human and machine. Importantly, some systems have functionalities that can provide explanations of how they arrived at a recommendation. Explanations have long been an essential and valued feature of AI-based systems because, by making the operation of the system transparent to the user, they can increase acceptance of the system and their trust in the advice provided (Hayes-Roth & Jacobstein, 1994). In their seminal article, Gregor & Benbasat (1999) provided a comprehensive review of empirical studies involving explanations for knowledge-based systems, including rule-based expert systems. Their article proposed an organizing framework and concluded that explanations, when suitably designed, led to improved performance and learning, and more positive perceptions of a system by users. We build on the Gregor and Benbasat framework as a foundation for our work. The interaction between users and AI-based system occurs in a context that includes the task the user is engaged in, their goals, and their broader environment. Providing explanations is one strategy that system designers can instantiate for such interactions in order to achieve a dedicated outcome (e.g., increase performance or trust by the user). In communication theory, discourse strategy relates to the "nature of the message" passing between the communicators in a given context (see Powers, 1995). In our context, these discourse strategies are instantiated in dedicated capabilities of the AI-based system. We use the term "capability" in the sense of Markus *et al.* (2002) to refer to the ability of the system to provide a certain functionality, e.g. a design feature can provide an explanation on how the AI-based system arrives at a recommendation.

NEW COMMUNICATIVE ACTION FRAMEWORK FOR DESGINING HUMAN – AI-BASED SYSTEM INTERACTION

The CA framework is part of a new design theory developed by the authors. The design theory proposes five discourse strategies with accompanying capabilities that can be used across a broad range of AI-based systems to achieve desired goals. It is important to note that the CA framework proposes discourse strategies from the AI perspective (i.e. the AI-based system can apply these strategies for the interaction with the human user). These discourse strategies have their origin in Habermas' theory of Communicative Action (Habermas, 1984). In the following, we introduce the five discourse strategies and give pointers on how we adapted the theory by Habermas (1984) to suit our human - AI-based system interaction context on the basis of prior theory and empirical work:

1) *Instrumental*. The goal of this discourse strategy is to allow for accurate, effective, and efficient performance. As long as the AI-based system gives accurate and intelligible directions to the human, then the system effectively achieves it goals. Justificatory theory is drawn from the field of human-computer interaction in general (e.g. Shneiderman & Plaisant, 2010). An example is the interface in a satellite navigation system in a car giving directions to the human driver.
2) *Influencing.* The goal is to achieve a course of action that benefits the AI-based system (i.e. its designers / owners) and that may or may not be of benefit to human users. Moreover, applying deceptions by the AI-based system is a possibility here as well. Justificatory theory includes theory of persuasion, e.g. the elaboration likelihood model (Petty & Cacioppo, 1986) and cognitive bias theory (Thaler & Sunstein, 2009). An example is a decision support system that influences users' investment decisions in a certain direction (Looney & Hardin, 2009).
3) *Dramatic.* The goal of this discourse strategy is to present a stylized representation of the AI-based system to the human user in order to impact user's affective perceptions. Justificatory theory for this strategy includes the computers are social actors paradigm (Nass *et al.*, 1994). An example is the human-like representation of an avatar for an recommender agent (Hess *et al.*, 2009).
4) *Normative.* The goal is to achieve a course of action that encourages or enforces compliance to societal, organizational or other norms. Justificatory theory includes social-norms theory (Schultz *et al.*, 2007). An example is a decision support system that utilizes weather forecasts and other information, including government policy, in planning winter road maintenance (Pisano *et al.*, 2004).
5) *Social.* The goal of this discourse strategy is to make arguments transparent and justifiable to allow all involved actors (i.e. AI-based system(s) and human user(s)) to reach understanding and achieve coordinated action. Justificatory theory includes Toulmin's model of argumentation (Toulmin *et al.*, 1984) and cognitive learning theory (Anderson, 1990). An example is a medical decision support system for managing hypertension that provides explanations based on varied sources (Shankar *et al.*, 2001).

In order to provide evidence of the validity of the proposed CA framework, we reviewed existing research on AI-based

systems to identify empirical support for the use of all five discourse strategies and overarching principles that connected their use. We found exemplar research across a number of sub-classes of AI-based systems instantiating at least one of the five discourse strategies (see examples in the list of the five strategies). There was also support for the overarching principles, with, for example, "explanations" in the social strategy leading to improved outcomes in many situations.

EXAMPLAR APPLICATION: APPLYING THE FRAMEWORK TO SOLVE A REAL-WORLD PROBLEM

We demonstrate the applicability of the proposed CA framework in a study concerned with suicide prevention at MetTrains. Suicides on railway systems are a serious problem world-wide. Suicides and attempted suicide impact not only the individuals involved, but also bystanders, railway staff and the travelling public. Research into measures to prevent suicide are ongoing. A meta-analysis by Havârneanu *et al.* (2015) indicates that measures with significant supporting evidence include: deterrence (platform screen doors, physical barriers, calming blue light, appropriate media reporting); detection (monitoring and detection system, surveillance unit); and response (pits between rails, staff training to approach people). MetTrains aims at continuing to reduce the incidence of suicide by train on its rail network and to this end has entered into a research collaboration to investigate how intelligent video surveillance can be employed to improve the detection of suspicious behaviour and allow effective response measures. The study had five phases:

1) Analysis and understanding of the work system into which the AI was to be introduced.
2) Analysis and understanding of the nature of the AI to be used, namely an AI classification system.
3) Use of the CA framework as a guide in a search to identify work on relevant interface capabilities, giving a "menu" of ideas for interface design.
4) A collaborative design workshop with stakeholders, using the "menu" as a base for idea stimulation but also allowing for new ideas from the users to emerge.
5) Design synthesis using the CA framework as an organizing device to show the features that the interface for a prototype Video Analytics (VAN) system would possess.

It should be noted that the design process we employed did not arise as a matter of course. A prior attempt by other project team members did not undertake phases 1 and 2 fully, or use the CA framework, and the first attempt at an interface design had a number of issues. For example, an indicator of the degree of risk of an alert showed "certain" as one end point of a continuum, an outcome which is not possible with this ML classification system and which is misleading to MetTrains staff.

Phase 1 – MetTrains Work Systems

Our case study organization, MetTrains, has a number of preventative measures already in place, including a high level of fencing of railway corridors, visible staff presence at many railway stations and a Security Control Centre (SCC) that coordinates communication and responses by station staff, police and ambulance when there is an alert. Analysis of incidents in the MetTrains database shows that the SCC systems and processes are already effective in preventing suicides. In cases where there is an alert of suspicious behaviour by staff or members of the public, responses such as approach, physical restraint and stopping of trains can be deployed. SCC personnel use closed-circuit television (CCTV) to monitor the situation in many cases. The systems appear to be working well. For example, in 2018 there were only 12 completed suicides and yet 103 cases of "prevention", where an individual attempted self-harm through contact with a train but was stopped in time to avoid injury. The conclusion from this part of the study was that a new system, at least in the prototype stage, would be an "add-on" to the main system, with the new system assisting by identifying serious risks, then passing processing control to the existing systems.

Phase 2 - Classification Systems

Our specific application case is intelligent machine surveillance, which is an example of ML used to classify human behaviour (suspicious or non-suspicious). Classification systems are an important type of ML, and also include applications such as fraud detection, or detection of objectionable content on the web (Martens & Provost, 2014). Often the suspicious events detected by surveillance systems will be very small in number (rare), compared with the number of normal events.

In the video surveillance context, human operators are often able to judge from the live CCTV if an event is truly suspicious once they are notified of it. However, they are poor at monitoring video feeds for long periods without their attention waning. In this case the AI-based system serves by giving a "tap on the shoulder", and the human is the decision maker. The AI-based system, however, should still provide pertinent information in a well-designed interface.

Phase 3 - Ideas Search

In the third phase, relevant literature was searched for prior empirical work that could inform the design of the interface for the proposed video surveillance system. The CA framework was used as a sensitizing device to locate prior interface capabilities that otherwise may have been missed.

Surprisingly, the extant literature on interfaces for surveillance systems is sparse. For intelligent video surveillance systems, Suss *et al.* (2015) note that work "seems to focus on technological advancements and largely ignores how automation will affect the human CCTV operator" (p. 1). An interface for an earthquake

surveillance system in California provides details such as the location and magnitude of the suspected earthquake, plus an indicator of the warning's "accuracy": e.g. 97% (Faulkner *et al.*, 2014). A system available commercially shows the CCTV image with the incident of interest plus a label: e.g. "presence in danger zone" (aitek, 2019).

Comparing prior work with the CA framework shows that interface designers all adopted instrumental strategies and that social strategies were employed to some degree. The social strategies included a basic form of explanation in the labelling of the type of suspicious behaviour that was the reason for the alert. Designers also included indicators of the systems performance, with reference to accuracy, sensitivity and true positives. In terms of Toulmin's model of argumentation, this information is a "qualifier" that reflects the degree of confidence in moving from grounds to a conclusion (Toulmin *et al.*, 1984).

Phase 4 – Collaborative Design

In this phase a workshop was held with staff at MetTrains, with an initial session discussing findings of phases 1 to 3, then a design session of 1.5 hours with a 'collaborative sketching' approach (Sangiorgi *et al.*, 2012). Seven staff members from the MetTrains SCC took part. They were receptive to the findings of phases 1 to 3, with a comment from the SCC manager "*you cannot treat AI like other IT*".

For the design session they were given a scenario in which the prototype system would be used and then the instructions on how to perform this session. Both of the two design groups chose to include basic informative content (following an instrumental discourse strategy), including location and time and the video capture. Both groups wanted the suspicious activity labelled: e.g., loitering (a simple form of explanation from the social strategy).

Phase 5 – Design Synthesis

A check was made of the use of the CA framework by working through it with a senior analyst to see if it was understandable and prompted any more ideas. The framework was understandable. The analyst reiterated there was no need for an influencing or normative discourse strategy. Moreover, there was no need for a dramatic discourse strategy, in the sense of personalization, except that loud "beeping" by the AI-based system was important when a detection alarm came in, as the environment the operators work in is very crowded and noisy. A feeling that the AI-based system is excited or alarmed would be good. It was acceptable to give the system a name, e.g. VAN, which suggests a human-type actor (Nass *et al.*, 1995)

A report of the workshop activities and the interface design has been passed back to MetTrains for checking and it has been decided to proceed with the prototype development based on the design produced.

DISCUSSION AND CONCLUSIONS

This work-in-progress study demonstrates the applicability of a new design framework for discourse strategies for AI-based systems. It first gives an overview of the proposed design framework, which has five design principles and four overarching principles, based on an adaption of Habermas's theory of communicative action for the human - AI-based system interaction context. Second, the paper shows the exemplar application of the CA framework in the development of an interface for a video surveillance system for suicide prevention at MetTrains. The MetTrains application case shows that the CA framework can be used as a sensitising device in an initial search for design ideas in prior literature. These design ideas can then form a base for a collaborative design exercise with stakeholders to produce prototype designs. Moreover, the framework can then be used as a checklist to work through a prototype design with stakeholders to ensure that no capabilities have been missed, before design synthesis occurs. The advantage of the CA framework is that it provides an opportunity to identify a wide range of design ideas that could be relevant for an AI-based system interface. Our case study showed that the prototype developed included capabilities not present in a prior attempt that did not use the CA framework.

Several avenues for further work exist. Apart from further refinement of the interface in the MetTrains case, the CA framework should be validated in other application areas. In addition, there is an opportunity to investigate the application and potential improvement of the CA framework with researchers as well as practitioners in further studies. Especially the understanding and resulting instantiation of the proposed design principles in actual AI-based systems by designers will be an interesting case to investigate how design knowledge from IS research can be adapted to real world problems.

ACKNOWLEDGMENTS

We acknowledge funding by the Australian Research Council and our industry partners and the contributions by staff at MetTrains and colleagues in the wider research project of which this study forms a part.

REFERENCES

1. aitek (2019) *AiVu-Smart Rail* [WWW document]. URL https://www.aitek.it/en/video-analytics-aivu-smart-rail/, accessed 26 August 2109.
2. Anderson, J. R. (1990) *Cognitive Psychology and Its Implications,* 3rd. W. H. Freeman, New York.
3. Brynjolfsson, E. & McAfee, A. (2016) *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton & Company, New York, London.
4. Faulkner, M., Cheng, M., Krause, A., Clayton, R., Heaton, T., Chandy, K. M., Kohler, M., Bunn, J., Guy, R., Liu, A. & Olson, M. (2014) Community sense and

response systems. *Communications of the ACM*, **57** (7), 66–75. doi: 10.1145/2622633.
5. Gönül, M. S., Önkal, D. & Lawrence, M. (2006) The Effects of Structural Characteristics of Explanations on Use of a DSS. *Decision Support Systems*, **42** (3), 1481–1493. doi: 10.1016/j.dss.2005.12.003.
6. Gregor, S. & Benbasat, I. (1999) Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, **23** (4), 497–530.
7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2019) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, **51** (5), 1–42. doi: 10.1145/3236009.
8. Gunning, D. (2018) *Defense Advanced Research Projects AgencyProgram Explainable Artificial Intelligence (XAI)* [WWW document]. URL https://www.darpa.mil/program/explainable-artificial-intelligence, accessed 30 May 2018.
9. Habermas, J. (1984) *The Theory of Communicative Action Reason and the Rationalization of Society. Vol 1*. Beacon Press, Boston.
10. Havârneanu, G. M., Burkhardt, J.-M. & Paran, F. (2015) A systematic review of the literature on safety measures to prevent railway suicides and trespassing accidents. *Accident; analysis and prevention*, **81**, 30–50. doi: 10.1016/j.aap.2015.04.012.
11. Hayes-Roth, F. & Jacobstein, N. (1994) The state of knowledge-based systems. *Communications of the ACM*, **37** (3), 26–39. doi: 10.1145/175247.175249.
12. Hess, T., Fuller, M. & Campell, D. (2009) Designing Interfaces with Social Presence: Using Vividness and Extraversion to Create Social Recommendation Agents. *Journal of the Association for Information Systems*, **10** (12), 889–919.
13. Knight, W. (2017) The Dark Secret at the Heart of AI. *MIT Technology Review*.
14. Looney, C. A. & Hardin, A. M. (2009) Decision Support for Retirement Portfolio Management: Overcoming Myopic Loss Aversion via Technology Design. *Management Science*, **55** (10), 1688–1703. doi: 10.1287/mnsc.1090.1052.
15. Markus, L. M., Majchrzak, A. & Les Gasser (2002) A Design Theory for Systems that Support Emergent Knowledge Processes. *MIS Quarterly*, **26**, 179–212.
16. Martens, D. & Provost, F. (2014) Explaining Data-Driven Document Classifications. *MIS Quarterly*, **38** (1), 73–99.
17. Mittelstadt, B., Russell, C. & Wachter, S. (2019) Explaining Explanations in AI. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. Unknown (ed.), pp. 279–288. ACM Press, New York, New York, USA.
18. Nass, C., Moon, Y., Fogg, B. J., Reeves, B. & Dryer, D.C. (1995) Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, **43** (2), 223–239. doi: 10.1006/ijhc.1995.1042.
19. Nass, C., Steuer, J. & Tauber, E. R. (1994) Computers are social actors. In: *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. Adelson, B., Dumais, S., Olson, J. (eds.), pp. 72–78. ACM Press, New York, New York, USA.
20. Petty, R. E. & Cacioppo, J. T. (1986) The Elaboration Likelihood Model of Persuasion. In: *Communication and Persuasion*, pp. 1–24. Springer New York, New York, NY.
21. Pisano, P. A., Stern, A. D., Mahoney III, W. P., Myers, W. L. & Burkheimer, D. (2004) Winter Road Maintenance Decision Support System Project: Overview and Status. In: *Sixth International Symposium on Snow Removal and Ice Control Technology*. Transportation Research Board (ed.), pp. 3–17.
22. Powers, J. H. (1995) On the intellectual structure of the human communication discipline. *Communication Education*, **44** (3), 191–222. doi: 10.1080/03634529509379012.
23. Sangiorgi, U. B., Beuvens, F. & Vanderdonckt, J. (2012) User interface design by collaborative sketching. In: *Proceedings of the Designing Interactive Systems Conference on - DIS '12*. Unknown (ed.), p. 378. ACM Press, New York, New York, USA.
24. Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. (2007) The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, **18** (5), 429–434. doi: 10.1111/j.1467-9280.2007.01917.x.
25. Shankar, R. D., Martins, S. B., Tu, S. W., Goldstein, M. K. & Musen, M. A. (2001) Building an explanation function for a hypertension decision-support system. *Studies in health technology and informatics*, **84** (Pt 1), 538–542.
26. Shneiderman, B. & Plaisant, C. (2010) *Designing the user interface: Strategies for effective human-computer interaction,* 5th ed. Addison-Wesley, Boston.
27. Suss, J., Vachon, F., Lafond, D. & Tremblay, S. (2015) Don't overlook the human! Applying the principles of cognitive systems engineering to the design of intelligent video surveillance systems. In: *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6. IEEE.
28. Thaler, R. H. & Sunstein, C. R. (2009) *Nudge: Improving decisions about health, wealth and happiness*. Penguin Books, London.
29. Toulmin, S., Rieke, R. D. & Janik, A. (1984) *An Introduction to Reasoning,* 2nd ed. Macmillan; Collier Macmillan Publishers, New York, London.
30. Weller, A. (2017) *Challenges for Transparency*.