Intrinsic, Dialogic, and Impact Measures of Success for Explainable AI

Jörg Cassens[1] , **Rebekah Wegener**[2]
[1]University of Hildesheim, 31141 Hildesheim, Germany
[2]Paris Lodron University, 5020 Salzburg, Austria
cassens@cs.uni-hildesheim.de, rebekah.wegener@sbg.ac.at

Abstract

This paper presents a brief overview of requirements for development and evaluation of human centred explainable systems. We propose three perspectives on evaluation models for explainable AI that include intrinsic measures, dialogic measures and impact measures. The paper outlines these different perspectives and looks at how the separation might be used for explanation evaluation bench marking and integration into design and development. We propose several avenues for future work.

1 Explanations

Explanations are foundational to social interaction [Lombrozo, 2006], and numerous different approaches to achieving explainability have been proposed recently [Adadi and Berrada, 2018; Arrieta *et al.*, 2019; Doran *et al.*, 2017].

Criticisms of current research trends include that "accounts of explanation typically define explanation (the product) rather than explaining (the process)" [Edwards *et al.*, 2019]. Another criticism is that explanations are currently largely seen as a relatively uniform and definable concept, and even systems that take user goals with explanation into account treat it largely on the system side of development [Biran and Cotton, 2017]. Despite this, a human centred [Ehsan and Riedl, 2020] perspective on explanation in artificial intelligence is not new [Shortliffe, 1976; Swartout, 1983; Schank, 1986; Leake, 1992, 1995; Mao and Benbasat, 2000]. For example, Gregor and Benbasat [1999] point out that different user groups have different explanation needs.

We have earlier construed contextualised explanations based on user goals [Sørmo *et al.*, 2005]. This has been used to integrate explanatory needs in the system design process [Roth-Berghofer and Cassens, 2005; Cassens and Kofod-Petersen, 2007]. However, we have represented explanation as a static object rather than a dialogic process. This includes the ability of the technical system to make use of explanations as well, at least as part of the theoretical model, even if not in practical applications.

In our understanding, both human and non-human actors in heterogeneous socio-technical systems (or socio-cognitive, [Noriega *et al.*, 2015]) can be senders and receivers of explanations [Cassens and Wegener, 2019]. For example, a human should be able to "explain away" recommendations made by a diagnostic system in order to enhance the future performance. While we currently focus on the opposite situation, e.g. an artificial actor explaining its choice of recommendations to the human user, frameworks for designing explanation-aware systems should be able to account for different flows of explanations, at least in principle and by extension.

In order to distinguish this from views that see the machine as only the explainer, not the explainee, we make use of the established term explanation awareness [Roth-Berghofer *et al.*, 2007; Roth-Berghofer and Richter, 2008]. Our working definition is as follows:

- **Internal View**: Explanation as *part of the reasoning process* itself.
  - *Example:* a recommender system can use domain knowledge to explain the absence or variation of feature values, e.g. relations between countries
- **External View**: giving explanations of the found solution, its application, or the reasoning process *to the other actors*
  - *Example:* the user tells said recommender system why he chooses an apartment in Norway despite the system suggesting one in Sweden

Semiotics and philosophy as well as the human and social sciences provide a rich basis for applications in explainable AI [Miller, 2018]. There is sufficient empirical and theoretical evidence that explanations are generated, communicated, understood and used in ways that are:

- **Dialogic,** as suggested e.g. by Leake Leake [1995],
- **Contextualised,** as required by e.g. Fraassen van Fraassen [1980], comprised of
  - *Context Awareness* (knowing the situation the system is in) and
  - *Context Sensitivity* (acting according to such situation) Kofod-Petersen and Aamodt [2006]; Kofod-Petersen and Cassens [2011]
- **Multimodal,** as argued for by e.g. Halliday Halliday [1978] and being
- **Construed by user interest,** as noted by e.g. Achinstein Achinstein [1983].

Given these foundations, can a semiotic model of explanation as a form of multi-modal dialogic language behaviour in context be used to generate contextually appropriate explanations by computational systems? There is an extensive body of research focusing on generating and using explanations in AI. Currently, what is lacking is:

1. A theory of the **dialogic process** rather than a monologic product
2. A **cohesive theory of explanation** that is:
   - *contextually appropriate* (e.g. fitting people, topic, mode and place),
   - *semantically appropriate* (e.g. recognised as an explanation)
   - *lexicogrammatically optimal* (best possible multimodal realisation)
3. A framework for integrating explanatory capabilities in the whole **software development life-cycle**, from requirements elicitation over design and implementation through to its use
4. A framework for **evaluation measures**.

We will focus on the last aspect in the remainder of this paper. Research in particular when it comes to measuring the actual effectiveness and efficiency of explanations given to users still seems fragmented. We propose to measure explainability along three lines of inquiry. **Intrinsic measures** deal with the question of whether the system at hand can generate explanations at all. **Dialogic measures** look at whether the system's output is seen as an explanation by the users. Finally, **impact measures** ask whether the explanation generated is of any use. These questions should help to elicit and formalise requirements for explanations as well as find ways to evaluate solutions that are operationalised sufficiently to enable making claims of explainability that can be tested against and to further comparisons between systems and iterations of systems.

Explanations are needed during the whole life cycle of applications, from initial requirements elicitation over design and development processes to using the final system. Therefore, it makes sense to look at frameworks for measuring efficiency and effectiveness of explanations in the context of whole development and life cycle management processes. While quality measurements for explanation could eventually enable a final system score (for benchmarking purposes [Zhan *et al.*, 2019]), development is a cycle and it is contextual, and the goal is to be able to build "better" systems through "better" development processes, where explanatory success is part of success metrics. Given existing requirements for transparency, such perspective on evaluating explanations can also be part of a regulatory framework for ethical AI [Cath, 2018; Coeckelbergh, 2020; Erdélyi and Goldsmith, 2018].

2 Evaluations

Within HCI, a plethora of different instantiations of human centred development processes exist (e.g. [Beyer and Holtzblatt, 1997; Carroll, 2000; Cooper *et al.*, 2014; De Ruyter and Aarts, 2010; Holtzblatt and Beyer, 2016], to name a few). We should consider principles and methods for (designing and evaluating) explainability as additions to existing tool kits, agnostic to their use in established design processes whenever possible (limited by different ontological commitments).

Evaluation is central to Human-Computer Interaction, or rather: evaluations are central since they typically form a cycle and cover a system at various stages. While (formative and summative) evaluations are a cornerstone for human centred design, "it is far from being a solved problem" [MacDonald and Atwood, 2013]. We are generally in need for evaluation processes that are suited for emerging types of applications [Poppe *et al.*, 2007] and for sustainable and responsible systems development [Remy *et al.*, 2018].

But even if current (usability) evaluation methods [Dumas and Salzman, 2006] may ultimately fall short in the context of XAI, they can at least inform first iterations of evaluation standards. In particular when used in combination with theories and models from other areas, such as linguistics [Cassens and Wegener, 2008; Halliday, 1978; Wegener *et al.*, 2008], psychology [Kaptelinin, 1996], the cognitive sciences [Keil and Wilson, 2000], or philosophy [Achinstein, 1983; van Fraassen, 1980].

In this short paper, we cannot explore these contributions in detail, but we will briefly outline a tripartite model for capturing explanatory effectiveness that includes:

- **Intrinsic measures:** measures that pertain to the ability of a system to generate explanations.
  Can the system generate explanations?
- **Dialogic measures:** measures that pertain to interaction between the system and its users.
  Does the system's output work as an explanation for its users?
- **Impact measures:** measures that pertain to the potential, anticipated or actual impact of explanations.
  Is the explanation generated of any use?

We have separated these measures because each of these three types of measures has different methods for testing and they cover distinct aspects of what "explanatory success" can mean. It is only by combining these different perspectives that we can get a full picture of the explanatory performance of a system and the explanations that are a part of that system. While we can think of more perspectives, it is important to keep in mind that quality measures have to have a well defined scope and they need to be, indeed, measurable [Carvalho *et al.*, 2017]. Furthermore, for them to be able to improve processes in practice, they need to be sufficiently simple to apply.

2.1 Intrinsic Measures

These measure the ability of the system to generate explanations, both generally for the given context of use, but specifically the transparency and interpretability of the system itself or of aspects of the system such as ML models and data used as well as algorithmic and other design choices.

If a system or parts of a system are not transparent then it is unlikely to perform well on either dialogic or impact measures. We can think of intrinsic measures as a baseline for explainable AI – it is a necessary, but not sufficient condition. From a design process perspective, we will need to look at which components are necessary for explanation generation [Roth-Berghofer and Cassens, 2005]. Evaluating, we might explore the structure, modality and semantic characteristics of the different explanations to ensure that they are optimised for the situation. There are different specific methods that might be useful for intrinsic measures.

2.2 Dialogic Measures

Here we look at the question of whether that which has been generated actually works as an explanation to the user, in various conditions, situations and contexts. Under investigation is the shared semiotic process of explanation generator and explanation consumer. Different methods are going to be useful for dialogic measures including user studies, reaction studies, experimental studies and qualitative and quantitative methods in general. Explanations are inherently dialogic, so we are always going to want to know who is requesting the explanation, who is providing the explanation and how and why they are providing it. Tracking the exchange of information itself is a way to evaluate because it lets us see the reaction to the explanation.

Trustworthy AI could be an outcome of systems that score highly on dialogic measures. This does not mean that trustworthy systems will score well on impact measures, indeed, human and non-human agents are quite prepared to trust a system that may have negative impacts on their wellbeing. Trust can be engendered through a dialogically well performing malicious system and this is what makes impact measures so essential.

2.3 Impact Measures

Impact measures look at whether providing explanations offers benefits over the use of the system itself. These can be used both on an individual level and for larger systems.

For example, on the individual level, we might consider an adaptive learning system that offers explanations to further the learning goal [Sørmo *et al.*, 2005] a user might have. While dialogic measures can be used to evaluate whether such an explanation can function as an explanation to the student, it would remain unclear whether the explanation did actually improve learning outcomes.

These measures also look at the impact that the system can have in the world. How can it impact decisions, diagnoses, legal and access outcomes? The impact measures examine the potential, anticipated or actual impact of the system and the ability of the system to explain these repercussions to users in context. Here the concept of contextual AI is important because as Ehsan and Riedl argue, "if we ignore the socially situated nature of our technical systems, we will only get a partial and unsatisfying picture" [Ehsan and Riedl, 2020]. A good model of context is crucial for evaluating explanatory success [Kofod-Petersen and Cassens, 2007; Wegener *et al.*, 2008]. Ethical AI would be the outcome of a system that scores highly on impact measures. We would of course aim for beneficial and equitable AI, but ethical is at least a good baseline outcome. Here we might expect to see methods such as impact studies and hypothetical, scenario and risk modelling. It would be beneficial to know what the anticipated consequences of the explanation are for everyone involved.

3 Related Work

Mohseni *et al.* [2018] argue that the interdisciplinary nature of explainable artificial intelligence (XAI) "poses challenges for identifying appropriate design and evaluation methodology and consolidating knowledge across efforts". At the same time, this interdisciplinary approach is essential to the success of XAI. We view our suggestion as a way to complement, further consolidate, and operationalise their classification system for different goals in XAI.

Hoffman *et al.* [2018] propose a process model of explaining and suggest measures that are applicable in the different phases of their conceptual model. This compliments our (more abstract) notions of dialogic and (to a lesser degree) impact measures, whereas we see our notion of intrinsic measures as a prerequisite for their model. Both models can be systematically combined, depending on the need for granularity and aspects covered. Mueller *et al.* [2021] present some helpful higher-level psychological considerations that can serve as general templates for effective explanations.

Sokol and Flach [2020] introduce fact sheets with an extensive list of properties for different explanatory methods. This is complimentary to our approach and could be used to select methods supporting the measures chosen. A survey by Carvalho *et al.* [2019] on interpretability in machine learning is orthogonal to our model, with their results being useful for operationalisation of the intrinsic (e.g. their comparison of different methods) and the dialogic measures (e.g. the notion of explanation properties).

4 Conclusion

We propose a tripartite perspective on explanation in intelligent systems that aligns with (iterative and contextual) design and development processes of systems such that there is space for formative and summative evaluations. While it enables a final system score (which we propose for benchmarking purposes [Zhan *et al.*, 2019]), development is a cycle and it is contextual, and the goal is to be able to build "better" systems, where explanatory success is part of success metrics.

We have previously discussed the potential for Ambient Intelligence to be useful for creating explainable AI [Cassens and Wegener, 2019], particularly on the architecture level and with regard to capabilities subsumed [De Ruyter and Aarts, 2010]. We propose that the core characteristics and general architecture of ambient intelligent systems make them a good framework for developing XAI and that AmI systems themselves have the potential to become explanatory agents that can be mediators between humans and other systems. The concept of mediating explanatory instances has also been explored in the context of virtual explanatory agents [Weitz *et al.*, 2020] or as a user-specific "memory" of explanations [Chaput *et al.*, 2021].

Development of such mediators, concentrating explanatory capabilities in specialised agents that are contextually embedded in our surroundings and have the potential for personalisation and anticipatory interaction, could greatly benefit from a cohesive framework for measuring explanatory success from different perspectives.

References

Peter Achinstein. *The Nature of Explanation*. Oxford University Press, Oxford, 1983.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *arXiv preprint: 1910.10045*, 2019.

Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017.

John M Carroll. *Making use: scenario-based design of human-computer interactions*. MIT press, 2000.

Rainara Maia Carvalho, Rossana Maria de Castro Andrade, Káthia Marçal de Oliveira, Ismayle de Sousa Santos, and Carla Ilane Moreira Bezerra. Quality characteristics and measures for human–computer interaction evaluation in ubiquitous systems. *Software Quality Journal*, 25(3):743–795, 2017.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Jörg Cassens and Anders Kofod-Petersen. Explanations and case-based reasoning in ambient intelligent systems. In David C. Wilson and Deepak Khemani, editors, *ICCBR-07 Workshop Proceedings*, pages 167–176, Belfast, Northern Ireland, 2007.

Jörg Cassens and Rebekah Wegener. Making use of abstract concepts – systemic-functional linguistics and ambient intelligence. In Max Bramer, editor, *Artificial Intelligence in Theory and Practice II – IFIP 20th World Computer Congress, IFIP AI Stream*, volume 276 of *IFIP*, pages 205–214, Milano, Italy, 2008. Springer.

Jörg Cassens and Rebekah Wegener. Ambient explanations: Ambient intelligence and explainable ai. In Ioannis Chatzigiannakis, Boris De Ruyter, and Irene Mavrommati, editors, *Proceedings of AmI 2019 – European Conference on Ambient Intelligence*, volume LNCS, Rome, Italy, November 2019. Springer.

Corinne Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges, 2018.

Rémy Chaput, Amélie Cordier, and Alain Mille. Explanation for humans, for machines, for human-machine interactions? In *WS Explainable Agency in Artificial Intelligence at AAAI 2021*, pages 145–152, 2021.

Mark Coeckelbergh. *AI ethics*. MIT Press, 2020.

Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. *About Face (fourth edition): the essentials of interaction design*. John Wiley & Sons, 2014.

Boris De Ruyter and Emile Aarts. Experience research: a methodology for developing human-centered interfaces. In *Handbook of ambient intelligence and smart environments*, pages 1039–1067. Springer, 2010.

Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint: 1710.00794*, 2017.

Joseph S. Dumas and Marilyn C. Salzman. Usability assessment methods. *Reviews of Human Factors and Ergonomics*, 2(1):109–140, 2006.

Brian J Edwards, Joseph J Williams, Dedre Gentner, and Tania Lombrozo. Explanation recruits comparison in a category-learning task. *Cognition*, 185:21–38, 2019.

Upol Ehsan and Mark O. Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. *arXiv preprint: 2002.01092*, 2020.

Olivia J. Erdélyi and Judy Goldsmith. Regulating artificial intelligence: Proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 95–101, New York, NY, USA, 2018. Association for Computing Machinery.

Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4):497–530, 1999.

Michael A.K. Halliday. *Language as a Social Semiotic: the social interpretation of language and meaning*. University Park Press, 1978.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint 1812.04608*, 2018.

Karen Holtzblatt and Hugh Beyer. *Contextual design: Design for life*. Morgan Kaufmann, 2016.

Viktor Kaptelinin. Activity theory: Implications for human-computer interaction. In Bonnie A. Nardi, editor, *Context and Consciousness*, pages 103–116. MIT Press, Cambridge, MA, 1996.

Frank C. Keil and Robert A. Wilson. Explaining explanation. In *Explanation and Cognition*, pages 1–18. Bradford Books, 2000.

Anders Kofod-Petersen and Agnar Aamodt. Contextualised ambient intelligence through case-based reasoning. In Thomas R. Roth-Berghofer, Mehmet H. Göker, and H. Altay Güvenir, editors, *Proceedings of the Eighth European Conference on Case-Based Reasoning (ECCBR 2006)*, volume 4106 of *LNCS*, pages 211–225, Berlin, September 2006. Springer.

Anders Kofod-Petersen and Jörg Cassens. Explanations and context in ambient intelligent systems. In Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, and Laure Vieu, editors, *Modeling and Using Context – CONTEXT 2007*, volume 4635 of *LNCS*, pages 303–316, Roskilde, Denmark, 2007. Springer.

Anders Kofod-Petersen and Jörg Cassens. Modelling with problem frames: Explanations and context in ambient intelligent systems. In Michael Beigl, Henning Christiansen, Thomas R. Roth Berghofer, Kenny R. Coventry, Anders Kofod-Petersen, and Hedda R. Schmidtke, editors, *Modeling and Using Context – Proceedings of CONTEXT 2011*, volume 6967 of *LNCS*, pages 145–158, Karsruhe, Germany, 2011. Springer.

David B. Leake. *Evaluating Explanations: A Content Theory*. Lawrence Erlbaum Associates, New York, 1992.

David B. Leake. Goal-based explanation evaluation. In *Goal-Driven Learning*, pages 251–285. MIT Press, Cambridge, 1995.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Craig M. MacDonald and Michael E. Atwood. Changing perspectives on evaluation in hci: Past, present, and future. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 1969–1978, New York, NY, USA, 2013. Association for Computing Machinery.

Ji-Ye Mao and Izak Benbasat. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Managment Information Systems*, 17(2):153–179, 2000.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv preprint: 1811.11839*, 2018.

Shane T. Mueller, Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. Principles of explanation in human-ai systems. In *WS Explainable Agency in Artificial Intelligence at AAAI 2021*, pages 153–162, 2021.

Pablo Noriega, Julian Padget, Harko Verhagen, and Mark D'Inverno. Towards a framework for socio-cognitive technical systems. In A. Ghose, N. Oren, P. Telang, and J. Thangarajah, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems X*, volume LNCS, pages 164–181. Springer, 2015.

Ronald Poppe, Rutger Rienks, and Betsy Dijk. Evaluating the future of hci: Challenges for the evaluation of emerging applications. volume LNCS 4451, pages 234–250, 01 2007.

Christian Remy, Oliver Bates, Jennifer Mankoff, and Adrian Friday. Evaluating hci research beyond usability. In *Extended Abstracts of the 2018 CHI Conference*, pages 1–4, 04 2018.

Thomas R. Roth-Berghofer and Jörg Cassens. Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In Héctor Muñoz-Avila and Francesco Ricci, editors, *Case Based Reasoning Research and Development – ICCBR 2005*, volume 3630 of *LNAI*, pages 451–464, Chicago, 2005. Springer.

Thomas Roth-Berghofer and Michael M Richter. On explanation. *Künstliche Intelligenz*, 22(2):5–7, 2008.

Thomas Roth-Berghofer, Stefan Schulz, David B Leake, and Daniel Bahls. Explanation-aware computing. *AI Magazine*, 28(4):122, 2007.

Roger C. Schank. *Explanation Patterns – Understanding Mechanically and Creatively*. Lawrence Erlbaum, New York, 1986.

Edward H Shortliffe. Computer-based medical consultations: Mycin. *New York*, 1976.

Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

William R. Swartout. What kind of expert should a system be? xplain: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21:285–325, 1983.

Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning – perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, October 2005.

Bas C. van Fraassen. *The Scientific Image*. Clarendon Press, Oxford, 1980.

Rebekah Wegener, Jörg Cassens, and David Butt. Start making sense: Systemic functional linguistics and ambient intelligence. *Revue d'Intelligence Artificielle*, 22(5):629–645, 2008.

Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, pages 1–12, 2020.

Jianfeng Zhan, Lei Wang, Wanling Gao, and Rui Ren. Benchcouncil's view on benchmarking ai and other emerging workloads. *arXiv preprint: 1912.00572*, 2019.