Let's Evaluate Explanations!

Miruna-Adriana Clinciu and Helen Hastie

Abstract: ***Transparency is an important factor for robots, autonomous systems and AI, if they are to be adopted into our lives and society at large. Explanations are one way to provide such transparency and natural language explanations are a clear and intuitive way to do this, helping users to understand what a robot or AI is doing and why. In this abstract, we highlight the importance of defining what makes a good explanation. Furthermore, we discuss evaluation methods for explanations by leveraging existing natural language generation evaluation metrics.***

Human-Robot Interaction (HRI) is a field of study dedicated to understanding, designing, and evaluating robotic systems, with the aim of creating a meaningful interaction between robots and humans. In recent years, robotic systems have increased in complexity and this has led to the need to explain their behaviour and reasoning, in order to better understand their capabilities and prevent errors. This aligns with the EPSRC Principles of Robotics, "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead, their machine nature should be transparent" (EPSRC, 2020).

Presently, robot behaviour can be perceived as providing too little information about the robot's intent and internal workings. This prohibits clear mental models of what it can and cannot do for the user but also this lack of transparency can inhibit progress from the developer's perspective (Wortham et al. (2017)). This, in turn, raises ethical and safety concerns. With regards to AI, EU GDPR introduced the "right to explanation" in the Article 22 "Automated individual decision-making, including profiling" GDPR in 2018. According to the above-mentioned regulations and principles, there is no doubt that we need a level of transparency and that this transparency may need to be communicated to the user. The importance of explanations for building trust and transparency in intelligent systems has been investigated by several researchers (Kulesza et al., 2012; Lim et al., 2009; Bussone et al., 2015) and previous work has shown that explanations can increase user understanding (Garcia et al., 2018) and trust in an intelligent system (Lim et al., 2009).

The question is how do we define what a good explanation is? Effective questioning (Wilen and Jr., 1986), a method of explanation in pedagogy, could help us to define a strategy for creating different types of explanations, but this is not sufficient. It is necessary to extract the main properties/attributes of an explanation in order to decide what makes a good or bad explanation. Zemla et al. (2017) consider that an explanation can have the following attributes: alternatives, articulation, complexity, desired complexity, evidence credibility, evidence relevance, expert, external coherence, generality, incompleteness, internal coherence, novelty, perceived expertise, perceived truth, possible explanation, principle consensus, prior knowledge, quality, requires explanation, scope and visualisation and according to Yuan et al. (2011), an explanation should be precise and concise. How do we know which of these factors are important and contribute the most to an effective explanation and how do they vary depending on the user and the context?

We consider that an intuitive medium to provide explanations is through natural language. There has been much work on natural language generation (NLG) evaluation (Hastie and Belz, 2014; Novikova et al., 2017) and we can potentially use these NLG measures to gauge the quality of an automatically generated explanation and even similarity to a 'gold standard' explanation using automatic measures from machine translation such as BLEU (Papineni et al, 2002) and ROUGE (Lin, 2004). In this abstract, we focus on four important properties of explanations that intersect between NLG and XAI namely: informativeness, readability, clarity, effectiveness.

Firstly, informativeness is linked with accuracy and adequacy and "targets the relevance and correctness of the output relative to the input specification" (Novikova et al., 2018). Secondly, readability can be measured using automatic objective measures but also human subjective evaluation. Automatic evaluation could be done by applying traditional readability indices that could be used to evaluate explanations e.g. Flesch Kincaid (Ease, 2009), FOG (Gunning, 1969). Human Evaluation for readability could be achieved by asking a target group to rate explanations for reading ease and comprehensibility (TAUS, 2014). Thirdly. according to Manishina (2016), important semantic formalisms are clarity and intuitiveness. Natural language explanations represent "support sentences", which are sentences that provide further information about the topic sentence through examples, reasons, or descriptions (McWhorter, 2016). Evaluating explanations in terms of clarity could focus on linguistic phenomena such as the misplaced or dangling modifiers, wordiness and redundancy and tense. Correct syntax is not the only factor to affect clarity and can also include many other factors, such as how to introduce concepts and ideas new to the user in a way that is appropriate to their knowledge and previous experience. For this, we can turn to the fields education and intelligent tutoring systems for inspiration (Graesser, 2016). Fourth with regards effectiveness, as mentioned by Tintarev and Masthoff (2007), effective explanations should help humans make good decisions. Effectiveness could be evaluated by calculating the difference of understanding in a model before and after providing the explanation and validity of any resulting decision. This could be achieved through gamifying a task to reward for good decisions

(Gkatzia et al., 2017) or comparing a user's understanding before and after an explanation (Garcia et al., 2018).

In conclusion, there is a clear need to define evaluation metrics for natural language explanations, in order to decide what makes a good or bad explanation and thus, in turn, increase transparency and avoid confusion and misunderstanding. Our current work is concerned with explaining causal Bayesian Networks where participants evaluate human explanations for graphical models, in terms of informativeness, clarity and effectiveness, taking inspiration from existing natural language generation evaluation metrics. Other properties of explanations, such as scrutability, satisfaction, persuasiveness, efficiency, soundness, coherence and understandability, will be taken into consideration for future research. It's clear that this is a multidisciplinary endeavour and factors from fields such as linguistics, NLP/NLG, cognitive science, psychology, pedagogy, as well as robotics and engineering will need to be considered.

References

[1] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. (2015). The role of explanations on trust and reliance in clinical decision support systems. In Proceedings of the 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015. Institute of Electrical and Electronics Engineers Inc., Piscataway, New Jersey, US, 160–169. https://doi.org/10.1109/ICHI.2015.26

[2] Martin Caminada *et al*. (2014). Scrutable plan enactment via argumentation and natural language generation. In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014, Vol. 2. International Foundation for Autonomous Agents and Multiagent Systems (IFAA-MAS), 1625–1626.

[3] GDPR. (2018). Article 22 EU GDPR "Automated individual decision-making, including profiling." http://www.privacy-regulation.eu/en/article-22-automated-individual-decision-making-including-profiling-GDPR.htm. Online; accessed 23 January 2020.

[4] Todd Kulesza *et al*. (2012). Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In Proceedings of the Conference on Human Factors in Computing Systems (CHI '12). Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678

[5] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. (2009). Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09). Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[6] Elena Manishina. (2016). Data-driven natural language generation using statistical machine translation and discriminative learning. Theses. Université d'Avignon. https://tel.archives-ouvertes.fr/tel-01398776

[7] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. (2018). RankME:Reliable Human Ratings for Natural Language Generation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, 72–78. https://doi.org/10.18653/v1/N18-2012

[8] The Engineering and Physical Sciences Research Council (EPSRC). (2020). Principles of robotics. https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/. Online; accessed 23 January 2020.

[9] Nava Tintarev. (2007). Explaining recommendations. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 4511 LNCS. 470–474.https://doi.org/10.1007/978-3-540-73078-1_67

[10] Nava Tintarev and Judith Masthoff. (2007). Effective Explanations of Recommendations: User-centered Design. In Proceedings of the 2007 ACM Conference on Recommender Systems (Minneapolis, MN, USA) (RecSys '07). ACM, New York, NY, USA, 153–156. https://doi.org/10.1145/1297231.1297259

[11] William W. Wilen and Ambrose A. Clegg Jr. (1986). Effective Questions and Questioning: A Research Review. Theory & Research in Social Education 14, 2 (1986), 153–161. https://doi.org/10.1080/00933104.1986. 10505518

[12] Robert H. Wortham, Andreas Theodorou, and Joanna J. Bryson. (2017). Robot transparency: Improving understanding of intelligent behaviour for designers and users. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 10454 LNAI. Springer Verlag, Berlin, Germany,274–289. https://doi.org/10.1007/978-3-319-64107-2_22

[13] Xiu Yu. (2017). A Brief Study on the Qualities of an Effective Sentence. Journal of Language Teaching and Research, 801. https://doi.org/10.17507/jltr.0804.21

[14] Changhe Yuan, Heejin Lim, and Tsai-Ching Lu. (2011). Most Relevant Explanation in Bayesian Networks.The Journal of Artificial IntelligenceResearch42, 1 (Sept. 2011), 309–352. http://dl.acm.org/citation.cfm?id=2208436.2208445

[15] Zemla, J. C. *et al.* (2017). Evaluating everyday explanations. Psychonomic Bulletin and Review, 24(5), 1488–1500. https://doi.org/10.3758/s13423-017-1258-z

[16] Francisco J. Chiyah Garcia *et al*. (2018). Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models through a Multimodal Interface. In Proceedings of the 11th International Conference of Natural Language Generation (INLG), Tilburg, The Netherlands.

[17] Helen Hastie and Anja Belz (2014). A Comparative Evaluation Framework for NLG in Interactive Systems. In Proceedings of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland, May 2014

[18] Kishore Papineni *et al*. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. DOI:https://doi.org/10.3115/1073083.1073135

[19] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches out (WAS 2004), (1), 25–26. Retrieved from papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85

[20] Kathleen T. McWhorter. (2016). Pathways: Scenarios for Sentence and Paragraph Writing, Books a la Carte Edition, 4th Edition

[21] Novikova, J. *et al*. (2018). Why We Need New Evaluation Metrics for NLG (pp. 2241–2252). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/d17-1238

[22] Ease, F. R. (2009). Flesch–Kincaid readability test. Reading, Vol. 70, pp. 8–10.

[23] Gunning, R. (1969). The fog index after twenty years. Journal of Business Communication, 6(2), 3–13. https://doi.org/10.1177/002194366900600202

[24] Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. International Journal of Artificial Intelligence in Education, 26(1), 124–132. https://doi.org/10.1007/s40593-015-0086-4

[25] Gkatzia, D., Lemon, O., & Rieser, V. (2017). Data-to-Text Generation Improves Decision-Making Under Uncertainty. IEEE Computational Intelligence Magazine, 12(3), 10–17. https://doi.org/10.1109/MCI.2017.2708998

[26] TAUS (2014). Best Practices on Readability Evaluation. [ONLINE] Available at: https://www.taus.net/academy/best-practices/evaluate-best-practices/best-practices-on-readability-evaluation. [Last Accessed 12th March 2020].