KI - Künstliche Intelligenz (2020) 34:193–198
https://doi.org/10.1007/s13218-020-00636-z

TECHNICAL CONTRIBUTION

Measuring the Quality of Explanations: The System Causability Scale (SCS)

Comparing Human and Machine Explanations

Andreas Holzinger[1,2] · André Carrington[3] · Heimo Müller[4]

Received: 3 October 2019 / Accepted: 9 January 2020 / Published online: 21 January 2020

Abstract
Recent success in Artificial Intelligence (AI) and Machine Learning (ML) allow problem solving automatically without any human intervention. Autonomous approaches can be very convenient. However, in certain domains, e.g., in the medical domain, it is necessary to enable a domain expert to understand, *why* an algorithm came up with a certain result. Consequently, the field of Explainable AI (xAI) rapidly gained interest worldwide in various domains, particularly in medicine. Explainable AI studies transparency and traceability of opaque AI/ML and there are already a huge variety of methods. For example with layer-wise relevance propagation relevant parts of inputs to, and representations in, a neural network which caused a result, can be highlighted. This is a first important step to ensure that end users, e.g., medical professionals, assume responsibility for decision making with AI/ML and of interest to professionals and regulators. Interactive ML adds the component of human expertise to AI/ML processes by enabling them to re-enact and retrace AI/ML results, e.g. let them check it for plausibility. This requires new human–AI interfaces for explainable AI. In order to build effective and efficient interactive human–AI interfaces we have to deal with the question of *how to evaluate the quality of explanations* given by an explainable AI system. In this paper we introduce our System Causability Scale to measure the quality of explanations. It is based on our notion of Causability (Holzinger et al. in Wiley Interdiscip Rev Data Min Knowl Discov 9(4), 2019) combined with concepts adapted from a widely-accepted usability scale.

Keywords System causability scale (SCS) · Explainable AI · Human–AI interfaces

✉ Andreas Holzinger
andreas.holzinger@human-centered.ai;
andreas.holzinger@medunigraz.at

André Carrington
acarrington@ohri.ca

Heimo Müller
heimo.mueller@medunigraz.at

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria

[2] xAI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

[3] Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada

[4] Diagnostic and Research Institute of Pathology, Medical University Graz, Graz, Austria

Abbreviations

AI	Artificial intelligence
aML	Automatic (or autonomous) machine learning
DL	Deep learning
FRT	Framingham Risk Tool
iML	Interactive machine learning
ML	Machine learning
SCS	System Causability Scale
SUS	System Usability Scale

1 Introduction

Artificial intelligence (AI) is an umbrella term for algorithms aiming at delivering task solving capabilities comparable to humans. A dominant sub-field is automatic (or autonomous) machine learning (aML) with the aim to develop software that can learn fully automatically from previous experience to make predictions based on new data. One currently very successful family of aML methods includes deep learning (DL), which is based on the concepts of neural networks, and the insight that the depth of such networks yields surprising capabilities.

Automatic approaches are present in daily practice of human society, supporting and enhancing our quality of life. A good example is the breakthrough achieved with DL [2] on the task of phonetic classification for automatic speech recognition. Actually, speech recognition was the first commercially successful application of DL [3]. Autonomous software is able today to conduct conversations with clients in call centers; Siri, Alexa and Cortana make suggestions to smartphone users. A further example is automatic game playing without human intervention [4]. Mastering the game of Go has a long tradition and is a good benchmark for progress in automatic approaches, because Go is hard for computers [5] because it is strategic, although games are a closed environment with clear rules and a large number of games can be simulated for big data.

Even in the medical domain, automatic approaches recently demonstrated impressive results: automatic image classification algorithms are on par with human experts or even outperforms them [6]; automatic detection of pulmonary nodules in tomography scans detected the tumoral formations missed by the same human experts who provided the test data [7]; neural networks outperformed a traditional segmentation methods [8], consequently, automatic deep learning approaches became quickly a method of choice for medical image analysis [9]

Undoubtedly, automatic approaches are well motivated for theoretical, practical and commercial reasons. Unfortunately, such approaches have also several disadvantages. They are resource consuming, require much engineering effort, need large amounts of training data ("big data"), but most of all they are often considered as black-box approaches which do not foster trust and acceptance and most of all responsibility. International concerns are raised on ethical, legal and moral aspects of developments of AI in the last years, particularly in the medical domain [10]. One example of such international effort is the Declaration of Montreal.[1]

Lacking transparency means that such approaches do not expose explicitly the decision process [11]. This is due to the fact that such models have no explicit declarative knowledge representation, hence they have difficulty in generating the required explanatory structures which considerably limits the achievement of their full potential [12].

Consequently, in the medical domain a human expert involved in the decision process can be beneficial yet mandatory [13]. However, the problem is that many algorithms, e.g. deep learning, are inherently opaque, which causes difficulties both for the developers of the algorithms, as well as for the human-in-the-loop.

Understanding the reasons behind predictions, queries and recommendations [14] is important for many reasons. Among the most important reasons is trust in the results which is improved by an explanatory interactive learning framework, where the algorithm is able to explain each step to the user and the user can interactively correct the explanation [15]. The advantage of this approach, called interactive machine learning (iML) [16], is to include the strengths of humans, in learning and explaining abstract concepts [17].

Current ML algorithms work asynchronously in connection with a human expert who is expected to help in data pre-processing (refer to [18] for a recent example of the importance of data quality). Also the human is expected to help in data interpretation - either before or after the learning algorithm. The human expert is supposed to be aware of the problem's context and to correctly evaluate specific data sets.

The iML-approaches can therefore be effective on problems with scarce and/or complex data sets, when aML methods become inefficient. Moreover, iML enables important mechanisms, including re-traceability, transparency and explainability, which are important characteristics for any future information system [19].

The efficiency and the effectiveness of explanations provided by ML and iML require further study [20]. One approach to the problem examines how people understand explanations from ML by qualitatively rating the effectiveness of three explanatory models [21, 22]. Another approach measures a proxy for utility such as simplicity [11, 23] or response time in an application [24]. Our contribution is to directly measure the user's perception of an explanation's utility, including cause aspects, by adapting a well-accepted approach in usability [25].

2 Causability and Explainability

2.1 Definitions

A statement s (see Fig. 1) is either be made by a human s_h or a machine s_m. $s = f(r, k, c)$ is a function with the following parameters:

r representations of an unknown (or unobserved) fact u_e related to an entity,

k pre-existing knowledge, which is for a machine embedded in an algorithm, or made up for human by explicit, implicit and tacit knowledge,

[1] https://www.montrealdeclaration-responsibleai.com.

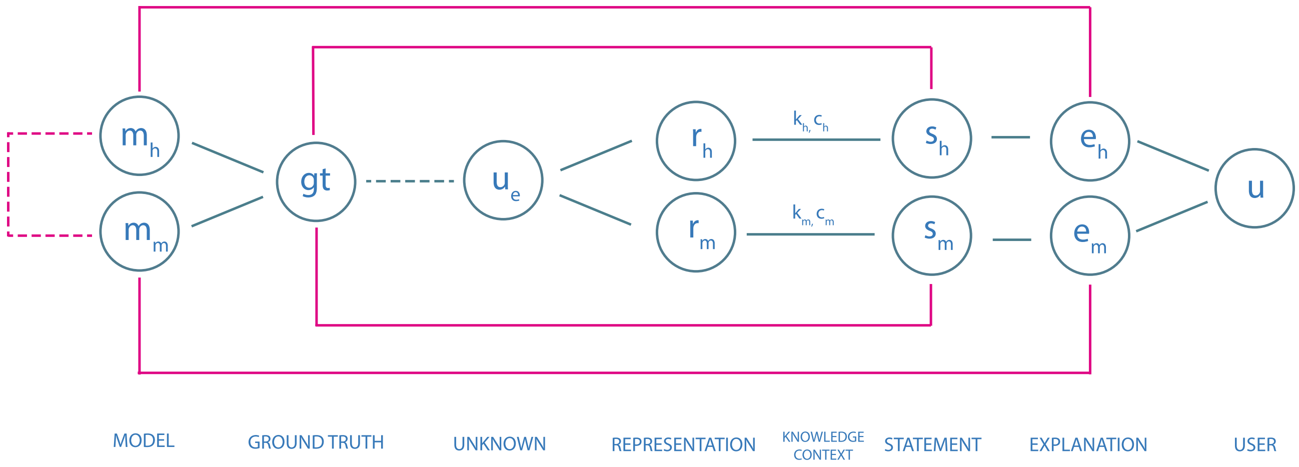

Fig. 1 The Process of Explanation. Explanations (e) by humans and machines (subscripts h and m) must be congruent with statements (s) and models (m) which in turn are based on the ground truth (gt). Statements are a function of representations (r), knowledge (k) and context (c)

c context, for a machine the technical runtime environment, and for humans the physical environment the decision was made (pragmatic dimension).

An unknown (or unobserved) fact u_e represents a ground truth *gt* that we try to model with machines m_m or as humans m_h. Unobserved, hidden or latent variables are found in the literature for Bayesian models [26], hidden Markov models [27] and methods like probabilistic latent component analysis [28].

The overall goal is, that a statement is congruent with the **ground truth** and the explanation of a statement highlights applied parts of the model.

3 Process of Explanation and the Importance of a Ground Truth

In an ideal world the human and machine statement are identical, $s_h = s_m$, and congruent with the ground truth, which is defined for machines and humans within the same, $m_h = m_m$ (a connection between them, see Fig. 1).

However, in the *real world* we face two problems:

(i) ground truth is not always well defined, especially when making a medical diagnosis; and
(ii) although human (scientific) models are often based on understanding causal mechanisms, today's successful machine models or algorithms are typically based on correlation or related concepts of similarity and distance.

The latter approach in ML is probabilistic in nature and is viewed as an intermediate step which can only provide a basis for further establishing causal models. When discussing the explainability of a machine statement we therefore propose to distinguish between

- Explainability, which in a technical sense highlights decision relevant parts of machine representations r_m and machine models m_m—i.e., parts which contributed to model accuracy in training, or to a specific prediction. It does not refer to a human model m_h.
- Causability [1] as the extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.

As causability is measured in terms of effectiveness, efficiency, satisfaction related to causal understanding and its transparency for a user, it refers to a human understandable model m_h. This is always possible for an explanation of a human statement, as the explanation is per se defined related to m_h. To measure the causability of an explanation e_m of a machine statement s_m either m_h has to be based on a causal model (which is not the case for most ML algorithms) or a mapping between m_m and m_h has to be defined.

4 Background

The System Usability Scale (SUS) has been in use for three decades and proved to be very efficient and necessary to rapidly determine the usability of a newly designed user interface. The SUS measures how usable a system's user-interface is, while our proposed System Causability Scale measures how useful explanations are and how usable the explanation interface is.

The SUS was created by John Brooke already in 1986 when working at the Digital Equipment Corporation

(DEC). 10 years later he published it as a book chapter [25] which received (as of 01.10.2019) 7949 citations on Google Scholar with an amazing trend upwards.

The success factor is *simplicity:* SUS consists of a 10 item questionnaire, each item having five response options for the end-users. Consequently, it provides a quick and dirty tool for measuring the usability, which proofed to be very reliable [29], and it is used for a wide variety of any products, not only user-interfaces [30].

When a SUS is used, participants are asked to score the following ten items with one of five responses that range from *strongly agree* to *strongly disagree*:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system

Interpreting SUS scores can be difficult and one big disadvantage is that the scores (since they are on a scale from 0 to 100) are often wrongly interpreted as percentages. The best way to interpret results involves normalizing the scores to produce a percentile ranking. Consequently, the participants scores for each question are converted to a new number, added together and then multiplied by 2.5 to convert the original scores of 0–40 to 0–100. Though the scores are 0–100, these are not percentages and should be considered only in terms of their percentile ranking.

Based on a lot of research, a SUS score above 68 would be considered above average and anything below 68 is below average, however the best way to interpret the results involves normalizing the scores to produce a percentile ranking.

A further disadvantage is that SUS has been assumed to be unidimensional. However, factor analysis of two independent SUS data sets reveals that the SUS actually has two factors Usable (8 items) and Learnable (2 items specifically, Items 4 and 10). These new scales have reasonable reliability (coefficient alpha of 0.91 and 0.70, respectively). They correlate highly with the overall SUS (r = 0.985 and 0.784, respectively) and correlate significantly with one another (r = 0.664), but at a low enough level to use as separate scales [31].

5 The System Causability Scale

In the following we propose our System Causability Scale (SCS) using the Likert scale similar to SUS. The Likert method [32] is widely used as a standard psychometric scale to measure human responses (see about the limitations in the conclusions). The purpose of our SCS is to *quickly* determine whether and to what extent an explainable user interface (human–AI interface), an explanation, or an explanation process itself is suitable for the intended purpose.

1. I found that the data included all relevant known causal factors with sufficient precision and granularity.
2. I understood the explanations within the context of my work.
3. I could change the level of detail on demand.
4. I did not need support to understand the explanations.
5. I found the explanations helped me to understand causality.
6. I was able to use the explanations with my knowledge base.
7. I did not find inconsistencies between explanations.
8. I think that most people would learn to understand the explanations very quickly.
9. I did not need more references in the explanations: e.g., medical guidelines, regulations.
10. I received the explanations in a timely and efficient manner.

As an illustration, SCS was applied by a medical doctor from the Ottawa Hospital (see the acknowledgement section) to the Framingham Risk Tool (FRT) [33]. FRT was selected as a classic example of a prediction model that is in use today.

FRT estimates the risk of coronary artery disease in 10 years for a patient without diabetes mellitus or clinically evident cardiovascular disease, and uses data from the Framingham Heart Study [34]. FRT includes the following input features: sex, age, total cholesterol smoking, HDL (high density lipoprotein) cholesterol, systolic blood pressure and hypertension treatment. The ratings for the SCS score are reported in Table 1.

6 Conclusions

The purpose of the System Causability Scale is to provide a simple and rapid evaluation tool to measure the quality of an explanation interface (human–AI interface) or an explanation process itself. We were inspired by the System

Table 1 Using SCS with the Framingham Model. Ratings are: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree

Question	Rating
01. Factors in data	3
02. Understood	5
03. Change detail level	5
04. Need teacher/support	5
05. Understanding causality	5
06. Use with knowledge	3
07. No inconsistencies	5
08. Learn to understand	3
09. Needs references	4
10. Efficient	5
$SCS = \sum_i Rating_i/50$	0.86

Usability Scale and the Framingham model which is often in use in daily routine. The limitations of the SCS is that Likert scales fall within the ordinal level of measurement, meaning that the response categories have a rank order. However, the intervals between values cannot be presumed equal (it is illegitimate to infer that the intensity of feeling between strongly disagree and disagree is equivalent to the intensity of feeling between other consecutive categories on the Likert scale). The legitimacy of assuming an interval scale for Likert-type categories is an important issue, because the appropriate descriptive and inferential statistics differ for ordinal and interval variables and if the wrong statistical technique is used, the researcher increases the chance of coming to the wrong conclusion [35]. We are convinced that our Systems Causability Scale is useful for the international machine learning research community. Currently we are working on an evaluation study with the application in the medical domain.

Acknowledgements Open access funding provided by Medical University of Graz. The authors declare that there are no conflict of interests. This work does not raise any ethical issues. The authors are grateful for feedback and input from Dr. Douglas Manuel, MD, MSc, FRCPC from the Ottawa Health Research Institute and for comments from the international research community. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 A reference model of explainable Artificial Intelligence for the Medical Domain.

References

1. Holzinger A, Langs G, Denk H, Zatloukal K, Mueller H (2019) Causability and explainability of AI in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 9(4)
2. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
3. Hinton G, Deng L, Dong Y, Dahl GE, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
4. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D (2017) Mastering the game of go without human knowledge. Nature 550(7676):354–359
5. Richards N, Moriarty DE, Miikkulainen R (1998) Evolving neural networks to play go. Appl Intell 8(1):85–96
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
7. Setio AAA, Traverso A, De Bel T, Berens MSN, van den Bogaard C, Cerello P, Chen H, Dou Q, Fantacci ME, Geurts B (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 42:1–13
8. Ghafoorian M, Karssemeijer N, Heskes T, van Uden IWM, Sanchez CI, Litjens G, de Leeuw F-E, van Ginneken B, Marchiori E, Platel B (2017) Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci Rep 7(1):5110
9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Snchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
10. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaney-Israni S, Goldenberg A (2019) Do no harm: a roadmap for responsible machine learning for health care. Nat Med 25(9):1337–1340
11. Carrington AM (2018) Kernel methods and measures for classification with transparency, interpretability and accuracy in health care. PhD thesis, The University of Waterloo
12. Bologna G, Hayashi Y (2017) Characterization of symbolic rules embedded in deep dimlp networks: a challenge to transparency of deep learning. J Artif Intell Soft Comput Res 7(4):265–286
13. Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 3(2):119–131
14. Valdez AC, Ziefle M, Verbert K, Felfernig A, Andreas H (2016) Recommender systems for health informatics: state-of-the-art and future perspectives. In: Andreas H (ed) Machine learning for health informatics, vol 9605. Lecture Notes in Artificial Intelligence LNAI. Springer, Berlin, pp 391–414
15. Teso S, Kersting K (2019) Explanatory interactive machine learning. In: AIES19 Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. AAAI
16. Holzinger A, Plass M, Kickmeier-Rust M, Holzinger K, Crian GC, Pintea C-M, Palade V (2019) Interactive machine learning: experimental evidence for the human in the algorithmic loop. Appl Intell 49(7):2401–2414

17. Holzinger A, Kickmeier-Rust M, Müller H (2019) Kandinsky patterns as IQ-test for machine learning. In International cross-domain conference for machine learning and knowledge extraction, Lecture Notes in Computer Science LNCS 11713. Springer, pp 1–14
18. Hassler AP, Menasalvas E, Garcia-Garcia FJ, Rodriguez-Manas L, Holzinger A (2019) Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. Springer/Nature BMC Med Inform Decis Making 19(1):33
19. Holzinger A, Kieseberg P, Weippl E, Tjoa AM (2018) Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In: Springer Lecture Notes in Computer Science LNCS 11015. Springer, pp 1–8
20. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
21. Chander A, Srinivasan R (2018) Evaluating explanations by cognitive value. In: International cross-domain conference for machine learning and knowledge extraction. Springer, Berlin, pp 314–328
22. Lou Y, Caruana R, Gehrke J (2012) Intelligible models for classification and regression. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 150–158
23. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1135–1144
24. Narayanan M, Chen E, He J, Kim B, Gershman S, Doshi-Velez F (2018) How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1802.00682
25. Brooke J (1996) SUS : a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL (eds) Usability evaluation in industry. Taylor and Francis, London, pp 189–194
26. Gelman A, Carlin JB, Stern HS, Dunson DB, Rubin DB (2013) Fundamentals of Bayesian data analysis: chapter 5 Hierarchical models. CRC Press, ISBN 978-1-58488-388
27. Fieguth P (2010) Statistical image processing and multidimensional modeling. Springer Science and Business Media, New York
28. Shashanka M, Raj B, Smaragdis P (2008) Probabilistic latent variable models as nonnegative factorizations. Comput Intell Neurosci
29. Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the system usability scale. Int J Hum Comput Interact 24(6):574–594
30. Holzinger A (2002) User-centered interface design for disabled and elderly people: First experiences with designing a patient communication system (PACOSY). In: Computer helping people with special needs, ICCHP 2002, Lecture Notes in Computer Science (LNCS 2398). Springer, pp 34–41
31. Lewis JR, Sauro J (2009) The factor structure of the system usability scale. In: International conference on human centered design, pp 94–103
32. Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140:1–55
33. Genest J, Frohlich J, Fodor G, McPherson R (2003) Recommendations for the management of dyslipidemia and the prevention of cardiovascular disease: summary of the 2003 update. CMAJ 169(9):921–924
34. Grundy SM, Pasternak R, Greenland P, Smith S, Fuster V (1999) Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the american heart association and the american college of cardiology. J Am Coll Cardiol 34(4):1348–1359
35. Jamieson S (2004) Likert scales: how to (ab)use them. Med Educ 38(12):1217–1218