Quantifying the Demand for Explainability

Thomas Weber, Heinrich Hußmann, Malin Eiband

Quantifying the Demand for Explainability

Thomas Weber thomas.weber@ifi.lmu.de[1][0000−0002−6894−605X], Heinrich Hußmann heinrich.hussmann@ifi.lmu.de[1][0000−0003−1709−7905], and Malin Eiband[1] malin.eiband@ifi.lmu.de

LMU Munich, Frauenlobstr. 7a, 80337 Munich, Germany
http://www.medien.ifi.lmu.de

Abstract. Software that uses Artificial Intelligence technology like Machine Learning is becoming ubiquitous with even more applications ahead. Yet, the very nature of these systems has made it very hard to understand how they operate, creating a demand for explanations. While many approaches have been and are being developed, it remains unclear how strong this demand is for different domains, application types, and user groups. To assess this, we introduce a novel survey scale to quantify the demand for explainability. We also apply this scale to an exemplary set of applications, novel and traditional, in surveys with 212 participants, showing that interest in explainability is high in general for intelligent systems but also traditional software. While this validates the heightened interest in explainability, it also reveals further questions, e.g. where we can find synergies or how intelligent systems require different explanations compare to traditional but equally complex software.

Keywords: explainable AI · XAI · survey · target group · use case

1 Introduction

Data-driven applications have become one of the driving factors of our modern society. They use technology like Machine Learning to process large amounts of data, enabling functionality that was previously very hard to achieve. Many of these allow us to let computers take over tasks that would otherwise take human intelligence and much time. They are often, therefore, also referred to as "intelligent systems" [28]. Their – seemingly non-intelligent – counterpart is software that does not use, e.g., Machine Learning but relies on a clear set of instructions and algorithms for its defined execution.

While traditional software had been the only type for many years, recent improvements, particularly in processing power and storage, have allowed intelligent systems to become ubiquitous also. Their "decisions" influence how we find and consume information, what we buy, how we communicate, etc. Their success is unlikely to stop, too, with up-and-coming applications like autonomous driving and personalized diagnostics heavily relying on their capabilities.

Naturally, intelligent systems thus can have a wide-ranging impact on our life. Understanding their effect on us and society, the ethical implications their

use has, etc., is, therefore, an essential ability for us all. The sheer complexity of these systems still impedes this, though.

Making these complex systems more accessible is one of the goals of explainable AI (XAI) [14, 15]. By providing explanations in various forms, this research branch hopes to allow people to understand how intelligent systems process data, make their decisions, and operate in general. While explanations of Machine Learning models often start from the technology, the human perspective is becoming more and more important [4, 16, 18, 25]. This includes methods for designing and building explainability for all sorts of target groups [16, 19, 25, 29].

One of the early questions in user-centred XAI must always be whether people do require and demand explanations of the systems they use. While often this demand appears logical, it is usually justified very qualitatively with the potential benefits of explanations. This, unfortunately, makes it very hard to quantify and compare this demand across different applications, groups, etc.

This paper contributes by introducing a survey scale in Sect. 3 for quantifying this *demand for explainability*. Not only does this allow for better justification of future explainability, but it also allows interesting comparisons. As an example and first application for this scale, we performed a high-level investigation into the difference between intelligent and traditional, non-intelligent systems.

The motivation for this comparison is that people have managed to use complex software systems for many years, often without explicit explanations. This begs the question of whether intelligent systems are, in this regard, fundamentally different to demand special attention. Depending on the differences and commonalities, it would also allow us to better find synergies between the specialized, emerging field of XAI and general, well-researched human-computer-interaction. We present the results of this first survey with 212 participants in Sect. 4 and subsequently discuss the implications.

2 Related Work

By design, AI or data-driven systems process volumes of data that are beyond human capabilities. While the concept of the learning process is fairly straightforward, understanding what an individual model has learned and how it comes to "decisions" is becoming more and more complex with the growing volume of input data. Explainable AI (XAI) is, therefore, a branch of research that aims to make these systems more explainable, interpretable, and understandable [15]. Over the years, people have developed many mechanisms, visualizations, and tools to provide explanations in some form on another [1, 3, 7, 14, 18–20, 22, 27].

However, many of the above methods to increase explainability can be very technology-focused making them not ideal for inexperienced end-users. As Miller et al. [23] highlight, this is exasperated by the fact that technology experts and the developers of these systems are often in charge of making them more understandable, which can lead to situations in which the end-users' demands are neglected.

There certainly are approaches to make AI more accessible [30, 2] and how their effect from a user-centred perspective [25]. Unfortunately, though, a study by Hase et al. [17] found that the effect which current explanations have is not as much as one would hope to begin with. Furthermore, inadequate explainability can also have an adverse effect where the users believe they understand the system even though they do not [13, 26]. Ehrlich et al. [9] also highlight the fact that with the uncertainty of data-driven systems, explanations may not always be beneficial, particularly in real-world scenarios when there may not be a correct choice, and situational judgement is required. A series of studies by Bunt et al. [5] furthermore indicates that explanations, while good, are not always considered worthwhile from the users' perspective. In fact, they found that their participants only desired explanations in about 7% of instances. At the same time, as Eiband et al. [10] explore, even placebo explanations can help.

All this suggests that explainability is not as simple as providing a nice explanation and all is well, but instead, some prior evaluation will be necessary to appropriately gauge when, how much, and what type of explanation is adequate.

3 Survey

The following section describes the survey and the scale to assess the demand for explainability as well as the aspects we took into account when designing it.

3.1 Demand for Explainability Scale

Since a survey of the literature yielded no readily available scale to measure the demand for explainability, we constructed our own.

To this end, we brainstormed an initial set of 30 questions related to explainability of software systems, which we narrowed down to 15 core questions based on the feedback of two experts for XAI and intelligent systems (see Tab. 1). All questions use five-point Likert scales.

After a pilot survey to eradicate comprehension issues, we used a first sample of 50 questionnaire answers for a factor analysis of those questions using R's *psych* package[1]. The factor analysis as a statistical method allowed us to determine which of our initial questions actually contribute to the intended topic and which are tangential [21].

The factor analysis revealed four factors (cf. Fig. 1) in our data with the loadings, as shown in Tab. 1. These values indicate how strongly the individual questions contribute to a common underlying topic. Visual analysis via a scree plot also supports the assumption that there are four factors. We concluded that the first group of questions actually pertains to the demand for explainability. The second may be considered to be about the system performance, the third appears to be about prior experience, and the last group covers the participants' perceived own expertise.

[1] https://www.rdocumentation.org/packages/psych/

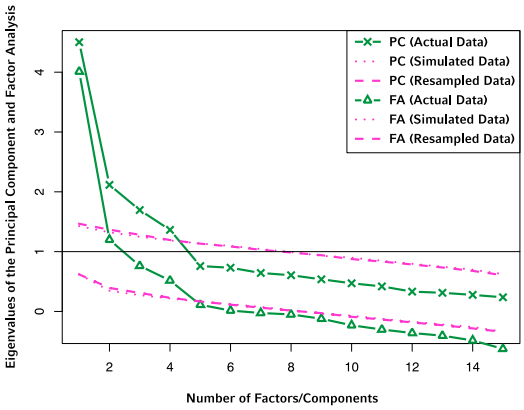

Fig. 1. The Scree plot, a method for visually identifying the number of factors during factor analysis, suggests four underlying factors.

We, therefore, considered only the first seven questions in the table as our scale throughout the remainder of this work. To see the scale work in practice, we then used it to gauge the demand for explainability for a number of software systems (as described below). We kept all 15 questions in, though, to re-run the factor analysis with a large data set, which yielded the same factors. This and a Cronbach's α of 0.88 for our scale indicate to us that the scale consistently measures what it is meant to.

3.2 Explainability of Different Types of Applications

To apply the scale in practice, we selected a number of software systems commonly found in the literature [6, 11, 12, 14, 19, 24, 29] for two comparisons.

First, we compared intelligent systems that people already use in their daily life to some that either still under research or used only in niche, experimental circumstances. With the explanations we provided for each of these systems, the participants could build an understanding of what an intelligent system is and thus had a point of reference and see intelligent systems, not as a vague concept but an everyday occurrence. For the systems not yet available, one would assume that the demand for explainability would be higher, already due to their unfamiliar nature, so it provided a good point of reference for comparison.

For the reasons mentioned at the beginning of this paper, we also chose to run a second comparison against more traditional software systems. The applications we used as examples in our surveys are, therefore, as listed in Tab 2.

For each application, we asked how frequently people use it and only applied the scale to those applications which the participant would use only at least infrequently. So, the number of applications for which each participant provided

Question	Loadings		
The systems needs to be explained more	0.833		
I would benefit from an explanation of how the system works	0.785		
The system should justify its decisions and its output	0.734		
It is important that people understand how this system works	0.719		
The system would be improved if it offered explanations for its behaviour	0.712		
Understanding how the system works is important to me	0.658		
I am interested in understanding the detailed internal workings of the system	0.608		
I am willing to trust the output of the system		0.636	
I would describe the system's behaviour as intelligent		0.485	
In my experience, the system works as intended		-0.546	
If the system did not work as intended, I would suffer negative consequences		-0.553	
I have been in a situation where the system behaved contrary to my expectations			0.690
I frequently experience situations in which I do not understand why the system behaves the way it does			0.686
I have an idea of what factors might influence the behaviour of the system			0.729
I believe I am capable of understanding how this system works			0.689

Table 1. Factor analysis indicated four factors in our 15 questions with the loadings as shown. Values of less than ±0.3 are omitted. We concluded that the first six questions refer to the *demand for explainability*.

feedback did vary slightly. To ensure a shared understanding of these applications, each was preceded by an example and a description which of its features are relevant for the survey and may warrant an explanation. We also used these descriptions to explicitly differentiate between intelligent and non-intelligent applications, particularly for those applications where the line starts to blur.

Besides this the survey also included a section for demographic and background information and the technology affinity scale by Edison and Geissler [8]. For each application, we also had open, free-text questions ask whether participants saw specific issues that require explanations in order to have additional qualitative feedback to the quantitative measure of the scale.

We disseminated each survey, comparison to future and traditional applications, online over a two-week period each to a diverse group of students, employees, and alumni of our institution.

4 Results

In total, 96 people completed the first survey comparing current and future intelligent systems, and another 116 contributed to the comparison to traditional software. In the following, we will explore these results of our survey scale.

Current Intelligent Systems	Future Intelligent Systems	Traditional Software
Online search engine	Autonomous driving	Operating systems
Social media platforms	Predictive policing	Web browser
Multimedia platforms	robotics	Email software
Online shops	Personalized medicine	Office applications
Navigation		Image manipulation software

Table 2. The applications we used as examples for the different groups in our survey.

Based on the demographic data, our survey participants provide us a sample of 94 males, 112 females, and six participants that chose not to disclose their gender. They tend to be on the younger side (mean: 24.8, sd: 5.7), which also shows in the education, where half had at most a high school education and a quarter had received a Bachelors degree, and in their professional experience, where a third indicated that they had less than a year of work experience (mean: 3.1, sd: 4.7). The technology affinity of our participants, measured on a range from 1 to 5, was decent, with an average score of 3.5 (sd: 0.9).

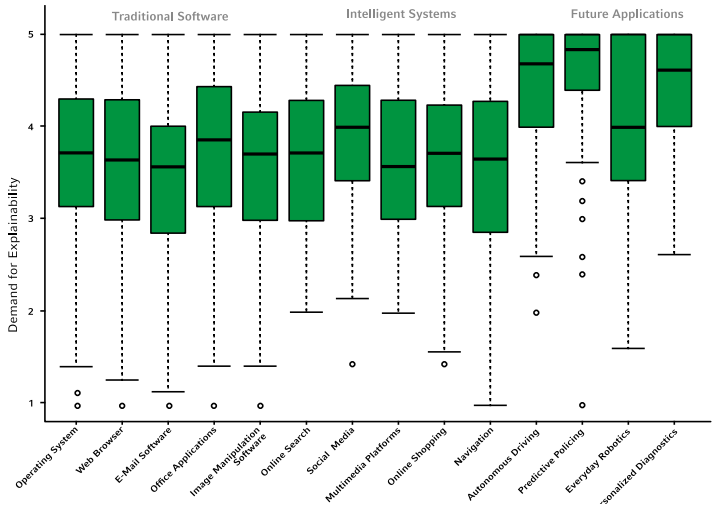

Fig. 2. The demand for explainability across the different applications.

Fig. 2 now shows the demand for explainability across our selected applications. Overall the demand is medium to high and very consistent within each application group. Between those groups, we can clearly see a difference as the demand for explainability is at a consistently medium level (mean: 3.6, sd: 0.7) for applications that the participants already use, while for the applications that are not yet widely available, on the other hand, the demand is overall

much higher (mean: 4.3, sd: 0.5). So we performed a hypothesis test to compare whether this difference between current and future applications is significant. With a Shapiro-Wilk-Test indicating a normal distribution of the data, the results of a subsequent ANOVA show a highly significant difference ($p < 0.001$) between these application types. Between traditional applications and available intelligent systems, we could not determine a significant difference ($p > 0.168$).

The quantification also allows us to better compare the demand for explainability between different participant groups. However, we could not determine any effect of age (corr. coeff: 0.055) or gender (t-test: $p > 0.62$) on the demand for explainability. Effects of education or professional experience also are negligible. The participants' technology affinity was ever so slightly correlated to the demand for explainability (correlation coefficient: 0.217) in general, although this results mostly from the traditional applications (correlation coefficient: 0.295) and not from the future applications (corr. coeff.: 0.091).

While the quantification alone is a good way for comparing two groups, it alone does not yet tell much about specific problems or solutions. This was very apparent in the qualitative list of issues that demand explainability which resulted from inductive coding of the qualitative answers. Concerns about data protection and privacy were raised 42 times for intelligent systems and only 5 times for traditional applications. Other issues were bias due to skewed data (29 intelligent / 0 traditional systems), security risks (9/6) or the effect of undetected failures (23/20).

Based on their feedback in the open questions, participants also disagreed on what format explanations should have across all applications, with some requesting detailed technical explanations while others explicitly stated that this type of explanation would not help them. The presentation of explanations, however, was not a separate question and thus only some participants divulged their preference. Consequently, we cannot make reliable estimates of how prevalent these opinions are for different applications.

5 Discussion

Based on the data outlined in the previous section, we will briefly discuss the meaning and implications of these results.

As a high-level summary, it can be said that the demand for explainability appears to be consistently high across the board. This high interest in explainability certainly validates the increased interest in XAI and human-centered methods in general. On the other hand, the fact that for existing intelligent applications the demand is not significantly higher, suggests that with all the interest in new technology, existing software is still not sufficiently explained. A possible explanation for this may be that even software systems that do not use Machine Learning or similar technology have become so complex that from the perspective of the average user they do not differ to much from data-driven applications. So, maybe research on explainability should focus more on inher-

ent complexity than on underlying technology. A future comparison to simpler systems may corroborate this.

One important factor that does play a role, though, is whether people already use an application, which leads to a slightly lower demand for explainability. This should come as no surprise, since the less they know about applications, the more information they need to understand it. People have gotten used to the software they use every day and are familiar with its capabilities and limitations.

The interesting point here is that many of these systems only have limited explanatory capabilities. This suggests that people are, in fact, capable of generating their own explanations even without explicit support. Whether these explanations are correct and sufficient is another question. Actively supporting the user still seems to be a worthwhile endeavor.

Considering the wide array of issue that our participant listed as in need of explanations, we can assume that not everyone needs the same explanations to be satisfied, though. The spectrum of issues is wide within single applications also. Furthermore, different participants gave different requirements for their explanations, strongly indicating that we will probably need multiple different explanations per application, target group, and use case, further pointing towards more user-centered approaches.

Which situation requires what type and level of explanation will need to be explored on an individual basis as our high-level assessment can only act as an early indicator. Clearly, there are many factors that have not yet been explored, like the influence of experience or personality. We, therefore, encourage future researchers and engineers to start not with the technology when building explanations but rather assess the demand quantitatively and qualitatively first and only then attempt to meet it.

6 Conclusion

In this paper, we introduced a survey scale to assess the demand for explainability in various software systems, domains and use cases. A quantification with such a test instrument is helpful for comparing systems, situations or user groups.

We tested this scale with an online survey where we recorded the general demand for explainability for various software systems, some of which use Machine Learning, some of which do not. This high-level evaluation validates the interest in XAI but it also points out the fact that intelligent systems are by no means the only software systems in need of better or more explanations.

Additional qualitative feedback from the survey also shows that there are some similarities in terms of what aspects need to be explained, but also some application-specific issues. Rather than building explanation form the technology side, this should encourage a user-centred perspective for explainability since there is no single solution that fits all situations or user groups. It also indicates that there might be strong synergies between more traditional Human-Computer-Interaction and XAI.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Amershi, S., Weld, D.S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S.T., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E.: Guidelines for human-ai interaction. In: Brewster, S.A., Fitzpatrick, G., Cox, A.L., Kostakos, V. (eds.) Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019. p. 3. ACM (2019)
3. Barbalau, A., Cosma, A., Ionescu, R.T., Popescu, M.: A generic and model-agnostic exemplar synthetization framework for explainable ai (2020)
4. Bohlender, D., Köhl, M.A.: Towards a characterization of explainable systems. CoRR **abs/1902.03096** (2019)
5. Bunt, A., Lount, M., Lauzon, C.: Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. In: Duarte, C., Carriço, L., Jorge, J.A., Oviatt, S.L., Gonçalves, D. (eds.) 17th International Conference on Intelligent User Interfaces, IUI 2012, Lisbon, Portugal, February 14-17, 2012. pp. 169–178. ACM (2012)
6. Cohen, I.G., Graver, H.: A doctor's touch: What big data in health care can teach us about predictive policing. SSRN Electronic Journal (2019)
7. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM **63**(1), 68–77 (Dec 2019)
8. Edison, S.W., Geissler, G.L.: Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. Journal of Targeting, Measurement and Analysis for Marketing **12**(2), 137 – 156 (11 2003)
9. Ehrlich, K., Kirk, S.E., Patterson, J.F., Rasmussen, J.C., Ross, S.I., Gruen, D.M.: Taking advice from intelligent systems: the double-edged sword of explanations. In: Pu, P., Pazzani, M.J., André, E., Riecken, D. (eds.) Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI 2011, Palo Alto, CA, USA, February 13-16, 2011. pp. 125–134. ACM (2011)
10. Eiband, M., Buschek, D., Kremer, A., Hussmann, H.: The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–6. CHI EA '19, Association for Computing Machinery, New York, NY, USA (2019)
11. Eiband, M., Völkel, S.T., Buschek, D., Cook, S., Hussmann, H.: When people and algorithms meet: user-reported problems in intelligent everyday applications. In: Fu, W., Pan, S., Brdiczka, O., Chau, P., Calvary, G. (eds.) Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019. pp. 96–106. ACM (2019)
12. Gade, K., Geyik, S.C., Kenthapadi, K., Mithal, V., Taly, A.: Explainable ai in industry. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 3203–3204. KDD '19, Association for Computing Machinery, New York, NY, USA (2019)
13. Gaviria, C., Corredor, J.A., Zuluaga-Rendón, Z.: "if it matters, I can explain it": Social desirability of knowledge increases the illusion of explanatory depth. In: Gunzelmann, G., Howes, A., Tenbrink, T., Davelaar, E.J. (eds.) Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017. cognitivesciencesociety.org (2017)

14. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., Holzinger, A.: Explainable ai: The new 42? In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 295–303. Springer International Publishing, Cham (2018)
15. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.: XAI - explainable artificial intelligence. Sci. Robotics **4**(37) (2019)
16. Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S.: A systematic method to understand requirements for explainable ai (xai) systems (2019)
17. Hase, P., Bansal, M.: Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? (2020)
18. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. CoRR **abs/1812.04608** (2018)
19. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? CoRR **abs/1712.09923** (2017)
20. Lundberg, S.M., Erion, G.G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.: Explainable AI for trees: From local explanations to global understanding. CoRR **abs/1905.04610** (2019)
21. Mair, P.: Factor Analysis, pp. 17–61. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-93177-7_2, https://doi.org/10.1007/978-3-319-93177-7_2
22. Melis, M., Demontis, A., Pintor, M., Sotgiu, A., Biggio, B.: secml: A python library for secure and explainable machine learning (2019)
23. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. CoRR **abs/1712.00547** (2017)
24. Mittelstadt, B.D., Floridi, L.: Transparent, explainable, and accountable AI for robotics. Sci. Robotics **2**(6) (2017)
25. Ribera, M., Lapedriza, À.: Can we do better explanations? A proposal of user-centered explainable AI. In: Trattner, C., Parra, D., Riche, N. (eds.) Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20, 2019. CEUR Workshop Proceedings, vol. 2327. CEUR-WS.org (2019)
26. Rozenblit, L., Keil, F.C.: The misunderstood limits of folk science: an illusion of explanatory depth. Cogn. Sci. **26**(5), 521–562 (2002)
27. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems pp. 1–21 (2020)
28. Völkel, S.T., Schneegass, C., Eiband, M., Buschek, D.: What is "intelligent" in intelligent user interfaces?: a meta-analysis of 25 years of IUI. In: Paternò, F., Oliver, N., Conati, C., Spano, L.D., Tintarev, N. (eds.) IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020. pp. 477–487. ACM (2020)
29. Wang, D., Yang, Q., Abdul, A.M., Lim, B.Y.: Designing theory-driven user-centric explainable AI. In: Brewster, S.A., Fitzpatrick, G., Cox, A.L., Kostakos, V. (eds.) Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019. p. 601. ACM (2019)
30. Wickramasinghe, C.S., Marino, D.L., Grandio, J., Manic, M.: Trustworthy AI development guidelines for human system interaction. In: 13th International Conference on Human System Interaction, HSI 2020, Tokyo, Japan, June 6-8, 2020. pp. 130–136. IEEE (2020)