Normative vs Pragmatic: Two Perspectives on the Design of Explanations in Intelligent Systems

Malin Eiband, Hanna Schneider, Daniel Buschek
LMU Munich, Munich, Germany
{malin.eiband, hanna.schneider, daniel.buschek}@ifi.lmu.de

ABSTRACT

This paper compares two main perspectives on explanations in intelligent systems: 1) A normative view, based on recent legislation and ethical considerations, which motivates detailed and comprehensive explanations of algorithms in intelligent systems. 2) A pragmatic view, motivated by benefits for usability and efficient use, achieved through better understanding of the system. We introduce and discuss design dimensions for explanations in intelligent systems and their desired realizations as motivated by these two perspectives. We conclude that while the normative view ensures a minimal standard as a "right to explanation", the pragmatic view is likely the more challenging perspective and will benefit the most from knowledge and research in HCI to ensure a usable integration of explanations into intelligent systems and to work on best practices to do so.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Explanations; Intelligent Systems; Transparency.

INTRODUCTION

Explaining how a system works and thus making its underlying reasoning transparent can contribute positively to user satisfaction and perceived control [8, 9, 14] as well as to overall trust in the system [11], and its decisions and recommendations [3, 13]. The legal obligation to make intelligent systems transparent – as enforced by European Union's General Data Protection Regulation [1] (GDPR) in May 2018 – is nevertheless strongly disputed. Integrating transparency is a complex challenge and there are no agreed upon methods and best practices to do so. Critics argue that such regulations will lead to deceleration of technical innovations (as many useful machine learning algorithms are not or not entirely explainable [16]) and deterioration of user experiences (as explanatory information can quickly clutter the interface or overwhelm users [7]).

We often trust human decision making without completely understanding the rationale behind it. Why do we not invest the same trust in AI calculations that consistently yield good results? In this position paper we analyze two arguments for transparency: a normative one emphasizing the right to receive explanations and a pragmatic one viewing transparency as a precondition for effective use. We illustrate how both perspectives differ and how they affect the design of explanations in intelligent systems.

THE NORMATIVE VIEW: A RIGHT TO EXPLANATION

> *"[Algorithmic] decisions that seriously affect individuals' capabilities must be constructed in ways that are comprehensible as well as contestable. If that is not possible, or, as long as this is not possible, such decisions are unlawful [...]"* [6]

A normative view on algorithmic transparency implies that intelligent systems may *only* be used if their underlying reasoning can be (adequately) explained to users. Following Hildebrandt's argumentation above, this would also concern cases in which intelligent systems might yield better results than non-intelligent ones – transparency is to be favored over efficiency and effectiveness out of ethical and legal reasoning. This view can also be found in the GDPR in Articles 13 to 15 that, together with Articles 21 and 22, express what has been called a "right to explanation" [5], granting access to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of [automated decision-making] for the data subject"[2]. But what does "meaningful information" signify and what are the consequences of this perspective when we want to design intelligent systems? Most of us do not fully understand even the workings of non-intelligent systems we interact with in everyday life, including some that may have a serious impact on our safety and well-being, such as cars or other means of transportation. Do we apply double standards or are there unique properties of intelligent systems that justify this scepticism? One possible answer is that in non-intelligent systems, no matter how complex they may be, we *theoretically* have the option to inform ourselves about their workings, in particular in cases in which

[1] ec.europa.eu/justice/data-protection/; accessed 27 September 2017.

[2] http://eur-lex.europa.eu/legal-content/EN/TXT/, accessed 15 December 2017.

ExSS '18, March 11, Tokyo, Japan.

the system does not react as expected. This option is currently not available in most intelligent systems, which brings up several interesting questions: Is the mere *option* to obtain an explanation about a system's workings more important than the *actual design* of this explanation (i.e., what is explained and how)? Does having this option alone already strengthen the trust in a system? This would imply that an explanation does not necessarily have to be usable nor seamlessly integrated into the interface or the workflow – most importantly, it should be *available* to users, and it should reflect the underlying algorithmic processing in detail and as comprehensively as possible.

THE PRAGMATIC VIEW: FOCUS ON USABILITY

> *"No, no! The adventures first, explanations take such a dreadful time."* [1]

From a pragmatic perspective, the current lack of transparency in intelligent systems hampers usability since users might not be able to comprehend algorithmic decision-making, resulting in misuse or even disuse of the system [12]. Explanations thus serve as a means to foster efficient and effective use of an intelligent system, and should be deployed wherever necessary to support users and their understanding of the system's workings. The mere option for explanations or the right to explanation would not suffice in this case, since a pragmatic solution might also ask for a minimum of cognitive load and a seamless integration of explanations into the interface and the workflow – excessive explanations would additionally hinder usability and interfere with the user experience. This perspective is challenging in practice, since designers have to find the sweet spot between several different requirements: *What* kind of information, and in *what detail*, is actually interesting and helpful to users in a particular situation or during a particular interaction? *How* can it be presented to the user without hampering usability? As text or visualization? If so, which wording or what kind of visualization is appropriate to not overwhelm users but still adequately reflect the complexity of the algorithm? To approach the design of explanations in intelligent systems from a pragmatic point of view, HCI research has brought forth exemplary prototypes [7, 17] one may consider for guidance, or design guidelines, such as Lim and Dey's intelligibility types [10]. However, best practices are still missing to date.

DESIGN DIMENSIONS

We describe several design dimensions to characterize possible explanations which might arise from either one of the two presented perspectives. Some of these dimensions, such as *Spatial Embedding* or *Temporal Embedding*, have been similarly presented in prior work, e.g., on system intelligibility [15] or meta user interfaces [2]. Table 1 presents an overview of these dimensions. The following sections introduce them in more detail, also pointing out connections between them.

Goal

The main *Goal* of the explanation summarizes the different motivations for the two perspectives:

The normative view aims to achieve a comprehensive and detailed understanding on the user's part – even if this takes a lot of time and effort (see *Level of Detail*). At the same time, it is not necessary that users go through explanations to use the system, the mere presence of the option for explanation might be enough for many users. In that sense, the normative view uses explanations also with the goal of creating general "background trust".

In contrast, the pragmatic view employs explanations to achieve a (possibly limited, non-comprehensive) level of understanding that facilitates usability and effective use of the system (see *Presentation*). Thus, it is necessary that users encounter explanations at some point before or during their main tasks with the system (see *Temporal Embedding*). To ensure this, systems may want to integrate explanations more closely (see *Spatial Embedding*) to achieve what we might call "foreground trust".

Foundation

The *Foundation* informs the content of the explanation (i.e., *what* to explain?).

The normative view may take into account an expert's mental model as a "gold standard" to cover all details of the underlying algorithm in a comprehensive, but still human-readable form.

In contrast, the pragmatic view also puts more emphasis on considering the users' mental models, for example to tailor explanations to particularly assess and address incorrect or incomplete aspects of these models [4].

Presentation

The *Presentation* dimension covers how the explanation is presented to the user.

To achieve a comprehensive detailed understanding, the normative view could employ almost any format, including videos, plots, interactive exploration and dedicated contact/help options, possibly even a "hotline" service.

In contrast, the pragmatic view aims for a presentation that facilitates a balance between explanation and the actual main UI elements. This might be achieved, for example, with markers/icons, details-on-demand techniques, textual or pictorial annotations, or modifications of layout and UI elements.

Level of Detail

The desired *Level of Detail* of the explanation also varies between the two perspectives:

The normative view favors a highly detailed explanation with the goal of comprehensive understanding of the intelligent system's underlying algorithms.

In contrast, the pragmatic view may favor a less detailed overview to facilitate a basic understanding. To do so efficiently, this view may focus on certain aspects and neglect others deemed less important. This focus could be informed by a user-centred design process (see *Foundation*).

Spatial Embedding

The *Spatial Embedding* describes how the explanation is integrated into the system's GUI overall.

Dimension	Normative Realization	Pragmatic Realization
Goal	understanding, background trust	usability, effective use, foreground trust
Foundation	expert mental model	symbiosis of expert and user mental models
Presentation	videos, plots, interactive exploration, contact/help options	markers, details-on-demand, UI elements and annotations
Level of Detail	high, comprehensive	overview, efficient
Spatial Embedding	separate view, “help page”	directly integrated into UI
Temporal Embedding	accessed before/after main tasks	interleaved with main tasks
Reference	underlying algorithms in general	specific content, e.g., a specific recommendation

Table 1. Design dimensions for explanations in intelligent systems and their desired realizations as motivated by the two perspectives.

The normative view motivates a detailed explanation which might thus not be embedded into the main GUI at all. Instead, systems could add a separate view, such as a “help page”.

In contrast, the pragmatic view is motivated to embed explanations directly into the GUIs used for the main tasks of the system. This dimension is thus strongly linked to the presentation choices (see *Presentation*).

Temporal Embedding

The *Temporal Embedding* describes how the explanation is integrated into the temporal workflow with the system.

The normative view motivates a detailed explanation which might thus not be embedded into the main task workflow at all. Instead, the user might optionally access it before or after the main task (e.g., on a separate page, see *Spatial Embedding*). Hence, once accessed, the full explanation is revealed at once.

In contrast, the pragmatic view is motivated to embed explanations directly into the workflow, for example using annotations or other details-on-demand within the main GUI views. This implies that the explanation is revealed gradually over the course of the user’s main tasks with the system.

Reference

The *Reference* dimension describes to which elements the explanations relate to primarily.

The normative view aims to reveal the underlying algorithms, yet may not be interested in doing so for specific cases that users encounter during their individual workflow.

In contrast, by integrating explanations more directly, the pragmatic view’s references for explanations are the specific cases encountered by the individual user during their interactions.

CONCLUSION

In this paper, we sketched out two perspectives on transparency in intelligent systems – a normative and a pragmatic view. The distinction between these two allows us to discuss different approaches to designing explanations. If one takes a normative standpoint, the mere option to receive explanations about an algorithm is critical and sufficient. Explanations need to be detailed enough to satisfy users’ needs for information. To avoid cluttering the interface, these detailed holistic explanations might be separated from the main interface, e.g., in a help function. If one takes a pragmatic standpoint, explanations detached from the interface and workflow are unlikely to be effective, as one can expect that very few users will make use of this option. The goal of the pragmatic approach is rather to integrate small bites of explanations into the interface to increase users’ understanding of the system slowly and effortlessly over time. It is the design of such well thought-through interface concepts that reveal the systems functioning during the interaction where HCI knowledge and research will be most needed and impactful.

That said, both perspectives are not to be regarded as mutually exclusive but can likely be combined appropriately. The normative perspective can then be regarded as “must have” and the right to receive explanations as a minimal standard, even if explanations are not integrated in a user-friendly fashion. Integrating explanations elegantly where they are interesting and useful for users will then be the challenge to work on and we invite HCI researchers to jointly work on this already now.

REFERENCES

1. Lewis Carroll. 2011. *Alice’s Adventures in Wonderland*. Broadview Press.
2. Joëlle Coutaz. 2006. Meta-User Interfaces for Ambient Spaces. In *International Workshop on Task Models and Diagrams for User Interface Design*. Springer, 1–15.
3. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The Effects of Transparency on Trust in and Acceptance of a Content-based Art Recommender. *User Modeling and User-Adapted Interaction* 18, 5 (20 Aug 2008), 455. DOI: http://dx.doi.org/10.1007/s11257-008-9051-3

4. Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. To appear in *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18)*.

5. Bryce Goodman and Seth Flaxman. 2016. European Union Regulations on Algorithmic Decision-making and a "Right to Explanation". *arXiv preprint arXiv:1606.08813* (2016).

6. Mireille Hildebrandt. 2016. The New Imbroglio. Living with Machine Algorithms. *Janssens, L.(ed.), The Art of Ethics in the Information Society. Mind you* (2016), 55–60.

7. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. DOI: http://dx.doi.org/10.1145/2678025.2701399

8. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/2207676.2207678

9. Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing (VLHCC '10)*. IEEE Computer Society, Washington, DC, USA, 41–48. DOI: http://dx.doi.org/10.1109/VLHCC.2010.15

10. Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp '09)*. ACM, New York, NY, USA, 195–204. DOI: http://dx.doi.org/10.1145/1620545.1620576

11. Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping Trust through Transparent Design: Theoretical and Experimental Guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*. Springer, 127–136.

12. Bonnie M. Muir. 1994. Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems. *Ergonomics* 37, 11 (1994), 1905–1922.

13. James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O'Donovan. 2015. Getting the Message?: A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 345–356. DOI: http://dx.doi.org/10.1145/2678025.2701406

14. Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW '07)*. IEEE Computer Society, Washington, DC, USA, 801–810. DOI: http://dx.doi.org/10.1109/ICDEW.2007.4401070

15. Jo Vermeulen. 2014. *Designing for Intelligibility and Control in Ubiquitous Computing Environments*. Ph.D. Dissertation.

16. Nick Wallace. 2017. EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence. (25 January 2017). Retrieved 15 December 2017 from http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm.

17. Rainer Wasinger, James Wallbank, Luiz Pizzato, Judy Kay, Bob Kummerfeld, Matthias Böhmer, and Antonio Krüger. 2013. *Scrutable User Models and Personalised Item Recommendation in Mobile Lifestyle Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 77–88. DOI:http://dx.doi.org/10.1007/978-3-642-38844-6_7