Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Malin Eiband
LMU Munich
Munich, Germany
malin.eiband@ifi.lmu.de

Sarah Theres Völkel
LMU Munich
Munich, Germany
sarah.voelkel@ifi.lmu.de

Daniel Buschek
LMU Munich
Munich, Germany
daniel.buschek@ifi.lmu.de

ABSTRACT

The rise of interactive intelligent systems has surfaced the need to make system reasoning and decision-making understandable to users through means such as *explanation facilities*. Apart from bringing significant technical challenges, the call to make such systems explainable, transparent and controllable may conflict with stakeholders' interests. For example, intelligent algorithms are often an inherent part of business models so that companies might be reluctant to disclose details on their inner workings. In this paper, we argue that as a consequence, this conflict might result in means for explanation, transparency and control that do not necessarily benefit users. Indeed, we even see a risk that the actual virtues of such means might be turned into *dark patterns*: user interfaces that purposefully deceive users for the benefit of *other* parties. We present and discuss such possible dark patterns of explainability, transparency and control building on dark UX design patterns by Grey et al. The resulting dark patterns serve as a thought-provoking addition to the greater discussion in this field.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Explainability; Explanations; Transparency; Dark Patterns; Interpretability; Intelligibility; User Control; Intelligent Systems.

ACM Reference Format:
Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019*. 6 pages.

1 INTRODUCTION

Intelligent systems that are empowered by advanced machine learning models have successfully been applied in closed contexts to well-structured tasks (e.g., object recognition, translations, board games) and often outperform humans in those. These advancements fostered the introduction of intelligent systems into more sensitive contexts of human life, like courts, personal finance or recruiting, with the promise to augment human decision-making in those.

However, the effectiveness of intelligent systems in sensitive contexts cannot always be measured in objective terms. Often they need to take soft factors, like safety, ethics and non-discrimination, into account. Their acceptance will greatly depend on their ability to make decisions and actions interpretable to its users and those affected by them. Introducing interpretability through *explanation facilities* [15] is widely discussed as an effective measure to support users in understanding intelligent systems [9, 24]. Yet, these measures are located at the intersection of potentially conflicting interests between decision-subjects, users, developers and company stakeholders [36].

First, companies may not see the benefit to invest in potentially costly processes to include explanations and control options for users *unless* they improve their expected revenues in some way. Second, creating suitable explanations of algorithmic reasoning presents a major technical challenges in itself that often requires abstraction from the algorithmic complexity [28, 29]. Furthermore, those systems are often integrated with critical business processes. Companies might be reluctant to disclose explanations that honestly describe their reasoning to the public as it might have an impact on their reputation or competitive advantage. Forcing companies to do so by law, like the *right to explanation* as part of the European Union *General Data Protection Regulation (GDPR)* [32], will most likely not result in meaningful explanations for users.

Therefore, we see a danger that means for algorithmic explanation, transparency and control might not always be designed by practitioners to *benefit* users. We even see a risk that users might consciously be deceived for the benefit of *other* parties. Such carefully crafted deceptive design solutions have gained notoriety in the UI design community as *dark patterns* [3].

In this paper, we extend the notion of prominent dark UX patterns [13] to algorithmic explanation, transparency and control. We discuss situations of opposing interests between the creator and receiver of algorithmic explanation, transparency and control means that could be potentially argued as questionable or unethical and contribute to the discussion about the role of design practitioners in this process.

IUI Workshops'19, March 20, 2019, Los Angeles, USA

2 BACKGROUND

2.1 Explanations in Intelligent Systems

Haynes et al. define intelligent systems as *"software programs designed to act autonomously and adaptively to achieve goals defined by their human developer or user"* [15]. Intelligent systems typically utilize a large knowledge data base and decision-making algorithms. Following Singh [31], a system is intelligent if users need to *"attribute cognitive concepts such as intentions and beliefs to it in order to characterize, understand, analyze, or predict its behavior"*.

Many of the intelligent systems developed today are based on increasingly complex and non-transparent machine learning models, which are difficult to understand for humans. However, sensitive contexts with potentially significant consequences often require some kind of human oversight and intervention. Yet, even intelligent systems in everyday contexts often confuse users [11]. For example, social network users are not aware that the news feed is algorithmically curated [6]. These insights result in ongoing research activities to improve the interpretability of those systems. *Interpretability* is the degree to which a human can understand the cause of a decision [26]. Interpretability can be achieved either by *transparency* of the model's inner workings and data, or *post-hoc explanations* that convey information about a (potentially) approximated cause – just like a human would explain [24].

Different stakeholders (e.g., creator, owner, operator, decision-subjects, examiner) of an intelligent system may require different means of interpretability [35]. Creators may demand *transparency* about the system's algorithms, while operators might be more interested how well the system's conceptual model fits their mental model (*global explanation*). Decision-subjects, on the other hand, may be interested in the factors influencing their individual decision (*local explanation*). This paper focuses on the interplay between owners of intelligent systems and decision-subjects using it.

Explanation facilities [15] are an important feature of usable intelligent systems. They may produce explanations in forms of textual representations, visualizations or references to similar cases [24]. The explanations provided may enable users to better understand why the system showed a certain behaviour and allow them to refine their mental models of the system. Following Tomsett [35] we define *explainability* as the level to which a system can provide clarification for the cause of its decision to its users.

Previous research work suggests that explanation facilities increase users' trust towards a system [23, 28] and user understanding [10, 18, 20]. However, how to present effective and usable explanations in intelligent systems is still a challenge that lacks best practices [22]. Due to the complexity of intelligent systems, explanations can easily overwhelm users or clutter the interface [18]. Studies by Bunt et al. [7] indicate that the costs of reading explanations may outweigh the perceived benefits of users. Moreover, some researchers warn that it may also be possible to gain users' trust with the provision of meaningless or misleading explanations [36]. This might leave users prone to manipulation and give rise to the emergence of dark patterns.

2.2 Dark Patterns

In general, a *design pattern* is defined as a proven and generalizing solution to a recurring design problem. It captures design insights in a formal and structured way and is intended to be reused by other practitioners [12]. Design patterns originate from architecture [1], but have been adopted in other fields such as software engineering [12], proxemic interaction [14], interface design [33], game design [37], and user experience design [13]. In contrast, an *anti pattern* refers to a solution that is commonly used although being considered ineffective and although another reusable and proven solution exists [17].

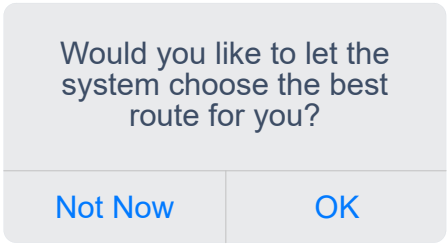

Figure 1: Exemplary interface for the *Restricted Dialogue* dark pattern. Users are not given a "No" option.

In 2010, Harry Brignull coined the term *dark pattern* [3] to describe *"a user interface that has been carefully crafted to trick users into doing things [...] with a solid understanding of human psychology, and they do not have the user's interests in mind"* [5]. He contrasts dark patterns to "honest" interfaces in terms of trading-off business revenue and user benefit [4]: while the latter put users first, the former deliberately deceive users to increase profit within the limits of law. Brignull [3] identified twelve different types of dark patterns and collects examples in his "hall of shame". Gray et al. [13] further clustered these dark patterns into five categories: *Nagging, Obstruction, Sneaking, Interface Interference* and *Forced Action.*

3 DARK PATTERNS OF EXPLAINABILITY, TRANSPARENCY AND CONTROL

What makes a pattern *dark* in the context of explainability, transparency and control? We see two general ways: the *phrasing* (of an explanation), and the way it is integrated and depicted in the *interface* (of explanation facilities). We build on the five categories of dark UX design patterns by Gray et al. [13] and apply them to the context of explainability, transparency, and user control, along with concrete examples (Table 1).

3.1 Nagging

Nagging is defined as a *"redirection of expected functionality that may persist over one or more interactions"* [13]. Transferred to the context of this paper, Nagging interweaves explanation and control with other, possibly hidden, functionality and thus forces users to do things they did not intend to do or interrupts them during their "actual" interaction.

3.1.1 Example 1: Restricted Dialogue. One example that Gray et al. present in their paper are pop-up dialogues that do not allow permanent dismissal. This could be easily transferred to our context: for example, an intelligent routing system could take control away from users with the tempting offer "Would you like to let the system

Dark Pattern by Gray et. al. [13]	**Transfer** to Explainability and Control	**Example Phrasings** of Explanation	**Example Interfaces** of Explanation Facilities
Nagging: *"redirection of expected functionality that may persist over one or more interactions"*	Interrupt users' desire for explanation and control	Restricted Dialogue	Hidden Interaction
Obstruction: *"making a process more difficult than it needs to be, with the intent of dissuading certain action(s)"*	Make users shun the effort to find and understand an explanation while interacting with explanation or control facilities	Information Overload, Nebulous Prioritization	Hidden Access, Nested Details, Hampered Selection
Sneaking: *"attempting to hide, disguise, or delay the divulging of information that is relevant to the user"*	Gain from user's interaction with explanation/control facilities through hidden functions	Explanation Marketing	Explanation Surveys
Interface Interference: *"manipulation of the user interface that privileges certain actions over others."*	Encourage explainability or control settings that are preferred by the system provider	Unfavorable Default	Competing Elements, Limited View
Forced Action *"Requiring the user to perform a certain action to access [...] certain functionality"*	Force users to perform an action before providing them with useful explanations or control options	Forced Data Exposure, Tit for Tat	Forced Dismissal

Table 1: Examples of dark patterns in the *phrasing* of explanations and the *interface* of explanation facilities. The examples are built upon the categorization by Gray et al. [13].

choose the best route for you?", where users can only select "Not now" or "OK", but have no "No" option (see Figure 1).

3.1.2 Example 2: Hidden Interaction. Nagging might include linking on-demand explanations with hidden advertisements: A click on "Why was this recommended to me?" on an ad could indeed open the explanation, but also the ad link (e.g., in two browser tabs).

3.2 Obstruction

Gray et al. define *Obstruction* in UX design as *"making a process more difficult than it needs to be, with the intent of dissuading certain action(s)"*. In the context of this paper, Obstruction makes it hard to get (useful) explanations about the system's decision-making and to control the algorithmic settings. Users thus might shun from the additional effort this takes and rather accept the system as is.

3.2.1 Example 1: Information Overload. Moreover, the use of very technical language to explain system behaviour and decision-making, or very lengthy explanations would most probably discourage users from reading the given information at all (see Figure 3. This might be comparable to what we currently see in end user licence agreements: the use of very technical language *and* a very lengthy presentation format results in users skipping the system prompt [2].

3.2.2 Example 2: Nebulous Prioritization. When explaining a decision or recommendation with a large number of influencing factors, the system might limit those factors by some notion of "importance" to not overwhelm the user. However, limiting factors requires a (potentially arbitrary) prioritization, which might be used to obfuscate sensitive factors, like family or relationship statuses. The explanation could be framed vaguely (e.g., "This recommendation is based on factors such as..." – i.e. not claiming to present all factors).

3.2.3 Example 3: Hidden Access. One way to obstruct the path to information could be to avoid "in-situ" links to explanations (e.g., offer no direct explanation button near a system recommendation). Instead, the option for explanation and control could be deeply hidden in the user profile and thus difficult to access.

3.2.4 Example 4: Nested Details. Similarly, the information *detail* could be distributed, for example nested in many links: When users want to have more than a superficial "This was shown in your feed, because you seem to be interested in fashion", they would have to take many steps to reach the level of detail that satisfies their information need.

3.2.5 Example 5: Hampered Selection. The system could also make activating explanations tedious for users by forcing them to do this for, say, every single category of product recommendation without giving a "select all" option. This could resemble the difficult cookie management practices seen today on many ad-financed websites. In another example setting, the information in an intelligent routing system could be spread along different sections of the recommended route and thus would have to be activated for each section separately.

3.3 Sneaking

The dark pattern of *Sneaking* is defined as *"attempting to hide, disguise, or delay the divulging of information that is relevant to the user"* [13]. Following this dark pattern, systems could use UI

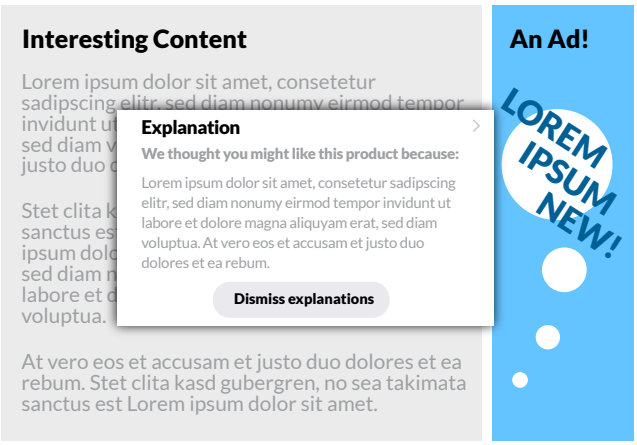

Figure 2: Exemplary interface for the *Limited View* dark pattern. Users are encouraged to dismiss explanations since they are layouted in a way that annoyingly covers the main content of the website.

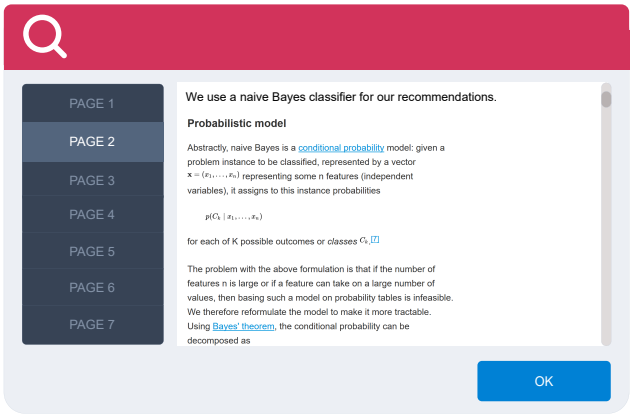

Figure 3: Exemplary interface for the *Information Overload* dark pattern. The given explanation is lengthy and uses technical language not suitable for non-experts (example article copied from Wikipedia).

elements for explainability and control, to sneak in information motivated by different intentions than interpretability.

3.3.1 Example 1: Explanation Marketing. For example, a web advertisement service could explain a particular ad by showing previously seen ads which the user had seemed to be interested in. Thus, the user's interest in an explanation is utilized to present multiple (potentially paid) advertisements. In a similar fashion, an online shop could use the opportunity of explaining product recommendations to promote further products. Also, ads might be directly integrated into the phrasing of explanations. For instance, an intelligent maps application might explain its routing decisions along the lines of "This route is recommended because it passes by the following stores you've visited in the past...".

3.3.2 Example 2: Explanation Surveys. Another approach might present an explanation and ask users for feedback in order to improve future explanations. This way, a company might enrich its user data and utilize it apart from explanation.

3.4 Interface Interference

Gray et al. [13] define this dark pattern as *"manipulation of the user interface that privileges certain actions over others."* In our context, this dark pattern privileges UI settings and user states that do *not* contribute to – or actively suppress – explainability, transparency, and user control.

3.4.1 Example 1: Unfavorable Default. For example, a dark pattern in this category could preselect a "hide explanations" option during the user onboarding in a financial robo-advisor system. This could be motivated to the user as "uncluttering" the dashboard or UI layout in general.

3.4.2 Example 2: Limited View. Explanations and control elements could also be layouted in a way that significantly reduces the space for the actual content or interferes with viewing it. This could encourage users to dismiss explanations to increase usability. Even simpler, links to an explanation might be presented in a barely visible manner. Figure 2 shows an example.

3.4.3 Example 3: Competing Elements. Further integration of explanations with the system's business model might involve, for instance, starting a count down timer upon opening an explanation for a booking recommendation to compete for the user's attention. This timer could indicate a guaranteed price or availability, thus putting pressure on the user to abandon the explanation view in order to continue with the booking process.

3.5 Forced Action

This dark pattern is defined as *"requiring the user to perform a certain action to access (or continue to access) certain functionality"* [13]. In our context, the user could be forced to perform an action that (also) dismisses functionality or information related to explainability, transparency and control.

3.5.1 Example 1: Forced Data Exposure. This dark pattern could be used to collect valuable user data under the pretext of explanation. The user might be forced to provide further personal information (e.g., social connections) before receiving personalized explanations. Otherwise, the user would be left off with a generic high-level explanation.

3.5.2 Example 2: Forced Dismissal. A user could be forced to dismiss an explanation pop-up in order to see the results of a request displayed underneath (e.g., during the investment process of a robo-advisor system). This dismissal might be interpreted as a permanent decision to no longer display any explanations.

3.5.3 Example 3: Tit for Tat. Regarding transparency, an e-commerce recommender system might force the user to first confirm an action (e.g., place an order) before it displays the factors that influenced

the recommendation. For instance, the system might proclaim that so far not enough data is available to explain its recommendation.

4 SUMMARY AND DISCUSSION

In this paper, we presented possible dark patterns of explanation, transparency and control of intelligent systems based on the categorization of dark UX design patters by Gray et al. [13]. We see the possibility that simple legal obligations for explanation might result in dark patterns rather than user benefits (e.g., similar to cookies settings on many ad-financed websites). Instead, with our work we intend to promote the on-going research on explainability as well as the discussion on explanation standards and their effects on users.

4.1 What Are the Consequences of Dark Patterns?

We see several possibly negative consequences of dark patterns in this context: Users might be annoyed and irritated by explanations, developing a negative attitude towards them. Examples include explanations presented in the *Nagging* patterns, which automatically open an advertisement along with the explanation; *Forced Action* patterns, which hinder the user to access desired results; or *Sneaking* patterns, which disguise advertisements as explanations. Similarly, users might lose interest in explanations when *Interface Interference* or *Obstruction* patterns are applied, which e.g., show long and tedious to read explanations. As a consequence, users might dismiss or disable explanations entirely.

On the other hand, users might not recognize explanations when they are hidden in profile settings. When users know that intelligent systems *must* provide explanations by law, the absence of explanations might mistakenly make users believe that the system does not use algorithmic decision-making. Hence, users might develop an incorrect understanding of algorithmic decision-making in general.

Furthermore, *Obstruction* patterns might lead to explanations which promote socially acceptable factors for algorithmic decision-making and withhold more critical or unethical ones. As a result, this might hinder the formation of correct mental models of the system's inner workings. Hence, users might not be able to critically reflect on the system's correctness and potential biases. As previous work in psychology suggests, users might accept *placebo* explanations without conscious attention as long as no additional effort is required from them [21]. When explanations use very technical language and are difficult to understand, users might simply skip them. This lack of knowledge and uncertainty about the underlying factors influencing the algorithm might lead to *algorithmic anxiety* [16].

4.2 Which Further Dark Patterns May Appear in this Context?

In this paper, we transferred the dark pattern categories by Gray et al. [13] to explainability and control of intelligent systems. However, there might be further patterns in this context. For example, we propose a pattern based on *Social Pressure* that uses information about other people – who are relevant to the user – in a way that is likely to be unknown or not endorsed by those people. For example, when Bob is shown an advertisement for diet products, explained by "Ask Alice about this", he might be annoyed with Alice without her knowledge. Similarly, Alice's boss might be recommended a lingerie shop that also "Alice might be interested in".

4.3 How Do Dark Patterns Affect Complex Ecosystems?

In this paper, we examined dark patterns which deceive decision-subjects who have means of directly interacting with the intelligent system. However, the ecosystem model of an intelligent system might be more complex and involve multiple stakeholders [35]. For example, in a financial decision-support context the system could ascertain the creditworthiness of a person (decision-subject), but only present an incontestable subset of reasons to the bank employee (operator) to not impact the reputation of the company (owner).

4.4 Can All Aspects of Dark Patterns Be Avoided?

Intelligent systems often use machine learning algorithms, which have hundreds of input variables. If all of these variables are explained, the explanation consists of a long list of text, which we identified as a dark pattern above. On the other hand, if they only show a subset of input variables for an explanation, this might bias the user's mental model, which is another dark pattern. Some explanations might be easier to understand for users than others. Hence, future studies have to evaluate which explanations are most helpful for users to understand the system.

4.5 How Can Dark Patterns Inform Research and Design?

In general, reflecting on dark patterns can be useful for HCI researchers and practitioners to learn how to do things properly by considering how not to do them. As a concrete use case, dark patterns can serve as a baseline for empirical studies to evaluate new design approaches: For example, a new explanation design could be compared against a placebo explanation – and not (only) against a version of the system with no explanation at all. Finally, dark patterns raise awareness that having *any* explanations is not sufficient. Instead, they motivate the HCI community to work on specific guidelines and standards for explanations to make sure that these actually support users in gaining awareness and understanding of algorithmic decision-making.

5 CONCLUSION

The prevalence of intelligent systems poses several challenges for HCI researchers and practitioners to support users to successfully interact with these systems. Explanations of how an intelligent system works can offer positive benefits for user satisfaction and control [19, 34], awareness of algorithmic decision making [27], as well as trust in the system [8, 25, 30]. Since 2018, companies are legally obliged to offer users a *right to explanation*, enshrined in the *General Data Protection Regulation* [32].

However, providers of intelligent systems might be reluctant to integrate explanations that disclose system reasoning to the public

in fear of a negative impact on their reputation or competitive advantage. Hence, legal obligations alone might not result in useful facilities for explanation and control for the end user.

In this paper, we have drawn on the notion of dark UX patterns [3] to outline questionable designs for explanation and control. These arise from explanation facilities that are not primarily designed with the users' benefits in mind, but purposely deceive users for the benefit of other parties.

In conclusion, we argue that while a legal right to explanation might be an acknowledgement of the necessity to support users in interacting with intelligent system, it is not sufficient for users nor our research community. By pointing to potential negative design outcomes in this paper, we hope to encourage researchers and practitioners in HCI and IUI communities to work towards specific guidelines and standards for "good" facilities for explanation, transparency and user control.

REFERENCES

[1] Christopher Alexander, Sara Ishikawa, Murray Silverstein, Max Jacobson, Ingrid Fiksdahl-King, and Shlomo Angel. 1977. *A Pattern Language: towns, buildings, construction.* Oxford University Press, Oxford, UK.

[2] Omri Ben-Shahar. 2009. The Myth of the "Opportunity to Read" in Contract Law. *European Review of Contract Law* 5, 1 (2009), 1–28.

[3] Harry Brignull. 2010. Dark Patterns. darkpatterns.org.

[4] Harry Brignull. 2011. Dark Patterns: Deception vs. Honesty in UI Design. https://alistapart.com/article/dark-patterns-deception-vs.-honesty-in-ui-design, accessed November 28, 2018.

[5] Harry Brignull. 2014. Dark Patterns: User Interfaces Designed to Trick People. http://talks.ui-patterns.com/videos/dark-patterns-user-interfaces-designed-to-trick-people, accessed November 28, 2018.

[6] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (jan 2017), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

[7] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI '12)*. ACM, New York, NY, USA, 169–178. https://doi.org/10.1145/2166966.2166996

[8] Henriette Cramer, Bob Wielinga, Satyan Ramlal, Vanessa Evers, Lloyd Rutledge, and Natalia Stash. 2009. The effects of transparency on perceived and actual competence of a content-based recommender. *CEUR Workshop Proceedings* 543 (2009), 1–10. https://doi.org/10.1007/s11257-008-9051-3

[9] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv e-prints* (Feb. 2017). https://arxiv.org/abs/1702.08608

[10] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (jun 2018), 1–37. https://doi.org/10.1145/3185517

[11] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When People and Algorithms Meet:Assessing User-reported Problems to Inform Support inIntelligent Everyday Applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA.

[12] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1994. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison Wesley, Boston, MA, USA.

[13] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 534, 14 pages. https://doi.org/10.1145/3173574.3174108

[14] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. Dark Patterns in Proxemic Interactions: A Critical Perspective. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[15] Steven R Haynes, Mark A Cohen, and Frank E Ritter. 2009. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies* 67, 1 (2009), 90–110. https://doi.org/10.1016/j.ijhcs.2008.09.008

[16] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 421, 12 pages. https://doi.org/10.1145/3173574.3173995

[17] Andrew Koenig. 1998. Patterns and Antipatterns. In *The Patterns Handbooks*, Linda Rising (Ed.). Cambridge University Press, New York, NY, USA, 383–389.

[18] 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. https://doi.org/10.1145/2678025.2701399

[19] 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. https://doi.org/10.1145/2207676.2207678

[20] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, New York, NY, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[21] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology* 36, 6 (1978), 635–642. http://dx.doi.org/10.1037/0022-3514.36.6.635

[22] Brian Y. Lim and Anind K. Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/2037373.2037399

[23] 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 27th International Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[24] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[25] 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*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Springer International Publishing, Cham, 127–136. https://doi.org/10.1007/978-3-319-41959-6_11

[26] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *CoRR* abs/1706.07269 (2017). arXiv:1706.07269 http://arxiv.org/abs/1706.07269

[27] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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 (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.

[30] 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. https://doi.org/10.1145/2678025.2701406

[31] Munindar P. Singh. 1994. *Multiagent systems.* Springer, Berlin, Heidelberg, Germany, 1–14. https://doi.org/10.1007/BFb0030532

[32] The European Parliament and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* (2016).

[33] Jenifer Tidwell. 2012. *Designing interfaces: Patterns for Effective Interaction Design.* O'Reilly Media, Inc., Sebastopol, Canada. arXiv:arXiv:gr-qc/9809069v1

[34] N. Tintarev and J. Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*. IEEE, New York, NY, UDA, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[35] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv e-prints* (June 2018). http://arxiv.org/abs/1806.07552

[36] Adrian Weller. 2017. Challenges for Transparency. *CoRR* (2017). http://arxiv.org/abs/1708.01870

[37] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark patterns in the design of games. In *Foundations of Digital Games 2013*.