From: Susan Von Struensee [redacted]
Sent: Saturday, June 19, 2021 3:45 PM
To: Comments
Subject: [EXTERNAL MESSAGE] RIN 3064-ZA24
Attachments: 2009.00246.pdf

Explainability Case Studies

Ben Zevenbergen Google benzevenbergen@google.com
Patrick Gage Kelley Google patrickgage@acm.org
Allison Woodruff Google woodruff@acm.org

ABSTRACT Explainability is one of the key ethical concepts in the design of AI systems. However, attempts to operationalize this concept thus far have tended to focus on approaches such as new software for model interpretability or guidelines with checklists. Rarely do existing tools and guidance incentivize the designers of AI systems to think critically and strategically about the role of explanations in their systems. We present a set of case studies of a hypothetical AI-enabled product, which serves as a pedagogical tool to empower product designers, developers, students, and educators to develop a holistic explainability strategy for their own products.

Explanations that come at the "moment of decision" are the current de-facto standard and best practice for AI explanations as implemented today. By contrast, in this work we take a broad view of explainability. This includes the exact, technical explanations that some AI systems provide at the moment of decision or inference—but also covers general introductions to an AI system, reasons given in moments of failure or error, descriptions of personalization and change over time, FAQ and help materials, design nudges and other less explicit information, results of audits and investigations, educational and public service campaigns, and any other support that helps users of a system understand the decisions made about them by AI. In this short paper, we present a set of case studies we have designed to support semi-structured, nuanced discussions of how explanations can be designed and deployed. Case studies are a particularly fruitful methodology to facilitate discussion and thinking about a range of variables that may influence outcomes, and they are often used in teaching technology ethics [2]. Instead of prescribing steps to take, our case study approach instills in participants a set of new perspectives that enables an approach to explainability beyond model interpretation and narrow in-the-moment notifications. The case studies highlight limitations of the status quo and encourage participants to explore a wider range of opportunities, challenges, and solutions than are commonly considered.

https://arxiv.org/pdf/2009.00246.pdf

Susan von Struensee

Explainability Case Studies

Ben Zevenbergen
Google
benzevenbergen@google.com

Patrick Gage Kelley
Google
patrickgage@acm.org

Allison Woodruff
Google
woodruff@acm.org

ABSTRACT

Explainability is one of the key ethical concepts in the design of AI systems. However, attempts to operationalize this concept thus far have tended to focus on approaches such as new software for model interpretability or guidelines with checklists. Rarely do existing tools and guidance incentivize the designers of AI systems to think critically and strategically about the role of explanations in their systems. We present a set of case studies of a hypothetical AI-enabled product, which serves as a pedagogical tool to empower product designers, developers, students, and educators to develop a holistic explainability strategy for their own products.

CCS CONCEPTS

• **Social and professional topics** → *Computing education*; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

explainability, accountability, AI, case studies, decision-making, ethics, machine learning, transparency

INTRODUCTION

Explainability has been highlighted as an important pillar of responsible AI practices [3, 5]. However, despite explainability being a focus of academic scholarship, technical development, industry guidelines,[1] and regulatory attention, best practices are not yet established for creating explanations that benefit individuals, communities, and expert audiences. Constructing good explanations for AI systems is a complex and largely untested design issue that does not yet lend itself to checklists but

[1] https://www.blog.google/technology/ai/ai-principles/, for example

rather calls for more open-ended exploration [11]. We aim to support AI designers, developers, educators, and others in the challenging task of considering how and where to deploy clear, understandable explanations that improve outcomes for individuals and society.

Explanations that come at the “moment of decision” are the current de-facto standard and best practice for AI explanations as implemented today. By contrast, in this work we take a broad view of explainability. This includes the exact, technical explanations that some AI systems provide at the moment of decision or inference—but also covers general introductions to an AI system, reasons given in moments of failure or error, descriptions of personalization and change over time, FAQ and help materials, design nudges and other less explicit information, results of audits and investigations, educational and public service campaigns, and any other support that helps users of a system understand the decisions made about them by AI.

In this short paper, we present a set of case studies we have designed to support semi-structured, nuanced discussions of how explanations can be designed and deployed. Case studies are a particularly fruitful methodology to facilitate discussion and thinking about a range of variables that may influence outcomes, and they are often used in teaching technology ethics [2]. Instead of prescribing steps to take, our case study approach instills in participants a set of new perspectives that enables an approach to explainability beyond model interpretation and narrow in-the-moment notifications. The case studies highlight limitations of the status quo and encourage participants to explore a wider range of opportunities, challenges, and solutions than are commonly considered.

RELATED WORK

Explainability as a concept has received much attention in academic, policy, and business literature [4, 7, 8, 10]. We present a snapshot of the literature due to space limitations.

Our case studies resonate with several ideas from academic literature. For example, a tiered system of transparency (through explanations) is highlighted in papers by Kaminski, Pasquale, and Edwards and Veale [1, 6, 9]. Our case studies also add to a growing body of resources for ethical AI. Tactical support for applying ethical AI ideas in practice is available in resources such as the Markkula Center Ethics in Technology Practice Framework and Toolkit,[2] the Omidyar Ethical OS Toolkit,[3] and the Princeton Dialogues on AI and Ethics Case Studies.[4]

Explainability also plays a role in checklists and toolkits that are largely aimed at having developers and designers build more ethical AI systems including Google’s People in AI Research Guidebook[5] and Responsible AI practices,[6] IBM’s AI Explainability 360,[7] and PwC’s Explainable AI.[8] Policymakers and regulators have also published guidelines and checklists focused on explainability. The EU’s High-Level Expert Group on Artificial Intelligence presents a brief selection of questions[9] and the British Information Commissioner’s Office (ICO) dedicates several publications to this topic.[10]

[2] https://www.scu.edu/ethics-in-technology-practice/

[3] https://ethicalos.org/

[4] https://aiethics.princeton.edu/case-studies/

[5] https://pair.withgoogle.com/

[6] https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability

[7] http://aix360.mybluemix.net/

[8] https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf

[9] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

[10] https://ico.org.uk/media/2616433/explaining-ai-decisions-part-2.pdf

THE CASE STUDIES

The explainability case studies presented in this paper are a pedagogical tool that are intended to be deliberated in a workshop setting. The cases intentionally include some questionable or problematic explainability practices. In each of five situations, workshop participants discuss how and why to improve the design of the AI systems and their explanations.

This approach takes participants out of their day-to-day context and places them into a hypothetical situation where an existing (though incomplete) explainability strategy is re-thought from the ground up. Using this approach, participants engage with ideas that underlie explainability best practices so they may then apply them in their own work.

The case studies revolve around an imagined high tech new car—The Model-U—with audio and visual sensors for identification of passengers, a personalized entertainment system, a reliable self-driving service, including taking complete driving control on highways, and advanced assistance in parking lots. Each of these features and systems are explored in the case studies.

Complete materials for using these case studies with a group of participants are available at: **https://arxiv.org/abs/2009.00246**

MATERIALS

We provide a readme which gives an overview of the case studies and guidance on how to run an in-person or virtual workshop. The slide deck contains an agenda for the activities, an introduction to the basic concepts, and an overview of the Model-U and its features. Each of the five situations has the written case, as well as discussion prompts to help participants work through the ethics and values questions the cases describe. Table 1 summarizes the main themes and a short description of each situation.

CONCLUSION

These materials are designed for audiences interested in the design and development of technology, including but not limited to practitioners, developers, user experience professionals, and undergraduate or graduate students. No specific background or expertise is required. We hope that people will find these discussions engaging and that they may apply the ideas in their own work designing and critiquing technology.

#	Situation	Themes
1	A family enters their newly purchased Model-U and are confronted with its identification system for the first time. One person is identified correctly and another is not, and the associated explanations cause user frustration.	• Superfluous content • Inappropriate timing • Explaining errors and uncertainty • Empowering user action
2	A driver merges onto the highway but is not ready to relinquish control of the Model-U to the self-driving system. The system's response makes an already complicated situation more stressful.	• Awareness of the user's context • Tone of explanation • Timeliness of content • Company response to user complaint about explanation
3	The car's entertainment system recommends music, but it is unclear how user feedback informs its choices. The driver realizes they don't know enough about how the system works.	• Providing meaningful feedback • Scarce attention • Incomplete mental models • Seeking appropriate moments for feedback
4	The Model-U is in a minor accident, and the driver receives a complex, formal explanation that is not meaningful to them. Investigation reveals the accident was caused by an adversarial attack, and the company's public response is not sufficiently reassuring.	• Varying end-user needs • Investigation of a high-profile failure • Public transparency • Communicating remote possibility of errors
5	The Model-U's traffic avoidance system leads to congestion in towns near highways. A local council organizes a stakeholder meeting which turns into a participatory design exercise. There is a gap between the information community members want so they can co-develop policy and what the company is willing or able to provide.	• Community participation and feedback • Providing information to the public • Responsiveness to diverse information requirements • Limitations on transparency

Table 1: A summary of the situations and themes in the five case studies.

REFERENCES

[1] Lilian Edwards and Michael Veale. 2017. Slave to the algorithm? Why a right to an explanation is probably not the remedy you are looking for. *Duke Law and Technology Review* 16 (2017).

[2] Casey Fiesler, Natalie Garrett, and Nathan Beard. 2020. What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 289–295. https://doi.org/10.1145/3328778.3366825

[3] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

[4] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[5] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1 (September 2019), 389–399.

[6] Margot E. Kaminski. 2019. The right to explanation, explained. *Berkeley Technology Law Journal* 34, 1 (2019).

[7] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[8] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 279–288. https://doi.org/10.1145/3287560.3287574

[9] Frank A. Pasquale. 2010. Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries. *Northwestern University Law Review* 104, 1 (2010), 105–174.

[10] Andrew D. Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham Law Review* 87 (2018), 1085–1139.

[11] Allison Woodruff. 2019. 10 things you should know about algorithmic fairness. *Interactions* 26, 4 (July 2019), 47–51. https://doi.org/10.1145/3328489