Answerability by Design in Sociotechnical Systems

DILARA KEKÜLLÜOĞLU, The University of Edinburgh, UK
LOUISE HATHERALL, The University of Edinburgh, UK
NAYHA SETHI, The University of Edinburgh, UK
TILLMANN VIERKANT, The University of Edinburgh, UK
SHANNON VALLOR, The University of Edinburgh, UK
NADIN KOKCIYAN, The University of Edinburgh, UK
MICHAEL ROVATSOS, The University of Edinburgh, UK

Sociotechnical systems (STSs) combine people and machines to take actions. Artificial intelligence (AI) enables STS to make increasingly autonomous decisions that impact human lives. Their reasoning processes still often remain unclear to people interacting with such systems, which may also harm people by making unjust decisions. There are no efficient means for people to challenge automated decisions and obtain proper restitution if necessary. On the other hand, organizations may be willing to provide more transparency about their decision-making process, but answering each of the questions people ask could be cumbersome. It is also not always clear who is qualified and accountable to answer to the people harmed by autonomous decisions. We argue that investigating the expectations of stakeholders is essential to create an efficient answerability framework. We propose a mediator agent that will bridge the gap between organizations that employ AI and people who were harmed by the automated decisions. Our approach helps the organizations to implement more answerable AI practices, and it also empowers people to ask for clarifications, request updates on actions as well as remedies through dialogues.

CCS Concepts: • **Human-centered computing** → **User studies**; • **General and reference** → **Design**; **Empirical studies**.

Additional Key Words and Phrases: artificial intelligence, responsible AI, AI harms, answerability, sociotechnical systems

ACM Reference Format:
Dilara Keküllüoğlu, Louise Hatherall, Nayha Sethi, Tillmann Vierkant, Shannon Vallor, Nadin Kokciyan, and Michael Rovatsos. 2023. Answerability by Design in Sociotechnical Systems. 1, 1 (September 2023), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ANSWERABLE SOCIOTECHNICAL SYSTEMS

Sociotechnical systems (STSs) bring together interconnecting systems that combine social and technological factors. Today's STSs are complex [3]: (i) they consist of many human and machine components, (ii) the outcomes of each of the components are observable whereas the internal workings may be unclear, (iii) there is a large number of interdependencies both internally and externally. This causes *responsibility gaps* where society expects and demands accountability for the actions but there are no clearly identifiable agents for moral responsibility [5, 12, 18]. Responsibility gaps are exacerbated by the fact that organizations which employ automated systems struggle to make their decision-making process transparent for their users who have questions, especially when their users are facing ethically significant harms. AI is a major contributor to the growing

Authors' addresses: Dilara Keküllüoğlu, d.kekulluoglu@ed.ac.uk, The University of Edinburgh, UK; Louise Hatherall, lhathera@ed.ac.uk, The University of Edinburgh, UK; Nayha Sethi, nayha.sethi@ed.ac.uk, The University of Edinburgh, UK; Tillmann Vierkant, t.vierkant@ed.ac.uk, The University of Edinburgh, UK; Shannon Vallor, svallor@ed.ac.uk, The University of Edinburgh, UK; Nadin Kokciyan, nadin.kokciyan@ed.ac.uk, The University of Edinburgh, UK; Michael Rovatsos, michael.rovatsos@ed.ac.uk, The University of Edinburgh, UK.

2023. XXXX-XXXX/2023/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

capability of some STS to issue automated decisions. Such decisions can be seen as insignificant and harmless in some application scenarios such as automated captioning. However, they become significant in high-impact situations such as medical diagnoses or loan applications.

Understanding how AI-based systems work is not trivial and people often struggle to find means to challenge their decisions. We believe that STSs using AI need to be *answerable* for the actions they take and the harms they inflict [5]. Being answerable also means being open to change, which brings organizations one step closer to implement *forward-looking responsibility* and to see responsibility as a virtue. Organizations can learn from previous inaccurate or harmful decisions and make better decisions in the future to prevent unintended results. Lyons et al. show that people–the recipients of the harms–value having a voice and a platform to challenge automated decisions [11]. Similarly, Cobbe et al. emphasize the importance of designing processes to review automated algorithms for improving accountability in STSs [4]. Hence, providing the users with the means to understand and challenge automated decisions is essential. This can also help STSs to become more trustworthy to be used in making informed decisions.

The General Data Protection Regulation (GDPR) [7] makes it clear that people have certain rights when they interact with AI systems. For example, Article 15 of the GDPR requires that individuals should be made aware of the existence of automated decision making and be given meaningful information about the logic involved. This has been clarified by the European Data Protection Board to include information to enable an individual to understand the reasons for the decision [6]. Article 22 also contains the right to contest decisions in certain circumstances, and seek human intervention on the part of the data controller where automated decision making processes are involved. Existing work shows that providing people with explanations about automated decisions increase their perceptions of informational fairness (i.e. to be given adequate information to understand about the process and decisions of the automated system) [16, 19]. Moreover, people value the ability to give new information to the system in order to update its actions [11].

Autonomous decisions have a substantial impact on our lives. To be socially responsible, organizations using AI-based systems should be able to explain the reasoning behind specific actions and provide their users a platform to understand their actions and challenge them if needed [7, 10]. Answerability is not only the explanation but also the capacity to be responsive to a wide range of context-specific expectations that answerable parties must meet to be considered responsible agents in society. For example, if there is an established harm, restitution is also necessary for answerable sociotechnical systems. We propose to investigate the stakeholder perceptions around autonomous decisions and the expectations of users, practitioners, and people who use autonomous decision-making tools in their professional lives. We follow with a real-case from a user who has been harmed by an AI system to show why we need tools to empower humans. We then introduce a mediator agent framework to establish a dialogue between the people who were harmed by an AI system's decisions and organizations using such an AI system. The mediator agent aims to provide timely responses while also enabling human assistance when needed. We also aim to expedite the process of complaints, identify the common fail points of a system to fix for future interactions, and allow organizations to be more responsible by allowing their users to appeal.

2 STAKEHOLDER PERCEPTIONS

Increased use of autonomous systems pose various legal and regulatory challenges [2, 13]. There remain gaps in determining legal responsibility for harms caused by autonomous systems [14], as well as in terms of how existing regulatory frameworks can be applied to the new technological advances ushered in by these systems [15]. Underpinning these legal and regulatory discussions is an emphasis that autonomous systems should be trustworthy. This is particularly true in high-stake contexts such as health [8]. Answerability is proposed here as a way of boosting the trustworthiness

of such systems. Understanding how to achieve this demands an exploration of the types of answers different stakeholders want and need in different contexts. However, there remain important epistemic gaps across the autonomous system landscape including who key stakeholders are, how they might be engaged with, and the nature of the relationships between diverse stakeholders when navigating the developing ecosystem. Understanding differences across contexts is also important: a patient who receives an incorrect diagnosis will need different answers than someone whose mortgage application was wrongly denied. Empirical data is vital to address these epistemic gaps, and to compare what answers are needed across differing contexts. Scoping conversations, interviews, and focus groups were carried out with a diverse range of stakeholders to identify the different answers people want and need in health, finance, and government. These findings will be useful both for the emerging regulatory landscape, and in the design of trustworthy autonomous systems.

3 HOW TO EMPOWER USERS LIKE KATHRYN?

Consider a real case where an AI system was used to compute a risk score for opioid addiction [17]. As a result, Kathryn–who suffers from endometriosis– had been denied access to crucial medication.

> During her hospital stay, Kathryn is given opioid medication because of her ongoing condition. One day a staffer informs her that she would no longer be receiving any kind of opioid with no further explanations. She later discovers that her pet's opioid medication was listed under her name, resulting in her high risk score.

In this case, the hospital was using an automated system [9] that calculates opioid addiction score according to features such as number of prescriptions, number of providers and pharmacies a person attended, and so on. Kathryn's risk calculation wrongly included the prescriptions of her sick dogs and caused the hospital to stop treatment. In this case, Kathryn should be able to get clear explanations from the hospital and appeal the decision quickly. However, it is not trivial to decide who should be answering to the Kathryn as the hospital is using a product of another company, and the state, pharmacies, other healthcare providers are sharing her data with the company. In this particular situation, Kathryn does not have any means to make anyone answer to her for the harm done, whether the answer she seeks is an apology, an explanation, a remedy, or restitution. We focus on building a mediator agent framework so that users like Kathryn could contest erroneous AI outputs or seek to understand the reasoning process better.

4 ANSWERABILITY THROUGH DIALOGUE

In order to make sociotechnical systems more answerable, we define the following three criteria for a mediator agent to help: (i) the system should aim to increase the user's understanding through its interactions, (ii) the harmful actions should be examined and addressed if necessary, (iii) the company should offer restitution efforts to compensate the harms where appropriate. We first explain the three dialogue stages of our approach: (1) Explanation, (2) Action Update, and (3) Remedy. The *explanation* stage will establish a common ground to exchange information and come to a shared understanding between the user and the mediator agent. The *action update* and *remedy* stages are more directed towards finding a solution for the harm that has been caused, if any. In all stages of the dialogue, the organization is involved in the process. For example, the mediator agent will contact appropriate people when it cannot find an adequate explanation, has missing information to provide an explanation, or the user does not accept the remedy offered.

Our approach gives answers to people who need help to understand the decision-making process of organizations, and it also provides an opportunity to companies to take responsibility for the actions of their systems. However, developing a practical solution to enable all these three stages

together is not trivial. We will now explain the components in detail and how the mediator agent could support end users to find the answers they need from organizations.

4.1 Stage 1: Explanation

The explanation phase provides a way for the user and the mediator agent to agree on a shared knowledge. This is a two-way communication where the mediator agent confirms its *assumptions* about the user and the information it already knows about the user. Throughout the explanation stage, the user should be able to give feedback on the reasons provided by the mediator agent. In other words, the user has the power to challenge the reasons they are provided. For example, if any of the reasons are based on outdated or incorrect information, the user can then correct this piece of information by guiding the agent, for example through an interactive chatbot. If the mediator agent cannot give reasons to particular actions by itself, it will request this information from the organization. This might involve using the various set of knowledge bases in the organizations, reaching out to a human representative of the organization, or even the developer teams who worked on the autonomous system who decided on the action. For this, the mediator agent will follow the *specification* provided by the organization, which includes information about what roles are available in the organization, and the specific members associated with such roles.

4.2 Stage 2: Action Update

The second part of our dialogue is the action update, which is triggered when the user requests the action to be updated. If the user requests the action update because the piece of information the action was based upon is either outdated or incorrect, the validity of the information can be checked. After this check the action can be updated by the organization if deemed possible. The dialogue then moves to the explanation step with the new action.

Action update might not be possible for every action. The organization might deem an update too costly. There can also be cases where updating the action would not help the user anymore (e.g. failed fraud detection). In these cases, the action update step will be skipped and the remedy part would be triggered. Whether the action update would happen or not is decided by the user and the representatives in the organization. The mediator agent will facilitate the conversation between the humans to reach a decision together.

4.3 Stage 3: Remedy

The last part of the dialogue is the remedy step, which is required when the user requests a remedy after seeing the explanations. Remedies can come in different forms such as monetary compensation, a promise to change the algorithm so other people would not be harmed in the same way in the future, and so on. A remedy can also be a combination of many remedy types; i.e. a remedy package. The mediator agent may offer a remedy package to the user who can then decide to accept or reject. Or both parties may have an ongoing conversation to agree on an acceptable remedy package for both sides. This will depend on the specific implementation of the framework.

5 EXAMPLE RUN: SUPPORTING KATHRYN

Kathryn did not know why the treatment was stopped and wanted to learn about the underlying reasons. We use this example with our implementation where Kathryn gets in contact with the hospital which employs our chatbot. For this example, we assume that prior to the conversation with the chatbot, it is established that Kathryn is a patient of the hospital and the action to stop treatment was based on a decision by an autonomous system. We will now focus on the following features: (1) *give* possible questions to ask from the knowledge base (2) *answer* the questions that were selected (3) *relay* the questions to the organization when they are not in the knowledge base

(4) *update* the action if possible (5) *offer* suitable restitution for the harms (6) *engage* the human representative when the user requests in any stage

5.1 Our Implementation

For the implementation of the chatbot, we use Rasa [1], an open-source Python library for conversational artificial intelligence. The domain knowledge can be given to the chatbot by training it with domain-specific examples. Rasa works on matching user input to defined *intents* and these intents have matching *actions* that will be taken by the chatbot. The chatbot will decide on the next action according to the policies given to it. These can be in form of rules (e.g., greetings will follow user greetings) or stories (e.g., a sequence of possible moves will be given). The actions can be static (e.g., say goodbye when the user says bye), or dynamic (e.g., find the capital of a country user gives from the internet).

We first define the possible intents and matching actions we need for this scenario. Some of the intents and actions can be seen in Table 1. We also give some example utterances given to train Rasa for specific intents. We created the rules that couple intents with actions, as well as stories that define example conversation flows to train our chatbot.

Table 1. Some examples of user intents, utterances, and resulting actions

Intent	Example Utterance	Action
ask_treatment_stop	"why did you stop my treatment?"	action_reason_treatment
why_at_risk	"what do you mean I have addiction risk?"	action_reason_risk
ask_score_system	"how did you calculate my risk score?"	action_explain_score
state_wrong_info	"this information is wrong"	action_wrong_info
accept_remedy	"i accept the offer."	action_accept_remedy
ask_human_assistance	"connect me to a representative."	action_connect_human

In our example, the chatbot starts with greetings and gives sample questions Kathryn can ask. The chatbot answers the selected questions using its knowledge base. When Kathryn states that the information around prescriptions is wrong, the chatbot triggers custom action to connect to a human representative. Kathryn asks for an action update and the hospital decides the start the treatment after verifying that incorrect information around prescriptions. Kathryn also asks for a remedy and accepts the offer given by the hospital.

6 CONCLUSION

To be socially responsible, organizations using AI-based systems should be able to explain the reasoning behind specific actions and provide their users a platform to understand their actions and challenge them if needed [7, 10]. If there is an established harm, restitution is also necessary for answerable sociotechnical systems. In this paper, we propose an agent-based framework for making sociotechnical systems (STSs) answerable, drawing on stakeholder perceptions gathered through scoping conversations, interviews, and focus groups. Currently, we define answerability in STSs, investigate stakeholder perspectives, design dialogue stages that are necessary. We also designed an ontology to define relations between actors. Our next steps will be expanding the scenarios we can support by abstraction of dialogue creation, running simulations with different user types and harms, and also conducting workshops with companies and developers to gain insight on their perceptions.

REFERENCES

[1] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181* (2017).

[2] Simon Chesterman. 2021. *We, the robots?* Cambridge University Press.

[3] Amit K. Chopra and Munindar P. SIngh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 48–53.

[4] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 598–609.

[5] Mark Coeckelbergh. 2020. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and eng. ethics* 26, 4 (2020), 2051–2068.

[6] European Commission. 2018. Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. Retrieved September 12, 2023 from https://ec.europa.eu/newsroom/article29/items/612053/en

[7] European Commission. 2018. Data protection in the EU. Retrieved May 3, 2023 from https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en

[8] Pouyan Esmaeilzadeh. 2020. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC medical informatics and decision making* 20, 1 (2020), 1–19.

[9] Bamboo Health. 2022. Yale New Haven Case Study. Retrieved September 5, 2023 from https://bamboohealth.com/wp-content/uploads/2022/10/Yale-New-Haven-Case-Study_Bamboo-Health.pdf

[10] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[11] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the appeal? Perceptions of review processes for algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[12] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6 (2004), 175–183.

[13] Lyria Bennett Moses. 2007. Recurring dilemmas: The law's race to keep up with technological change. *U. Ill. JL Tech. & Pol'y* (2007), 239.

[14] David Nersessian and Ruben Mancha. 2020. From automation to autonomy: legal and ethical responsibility gaps in artificial intelligence innovation. *Mich. Tech. L. Rev.* 27 (2020), 55.

[15] J Ordish. 2023. Large Language Models and software as a medical device. *Medicines and* 479 (2023).

[16] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. " There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022).

[17] Maia Szalavitz. 2021. The Pain Was Unbearable. So Why Did Doctors Turn Her Away? Retrieved February 26, 2023 from https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

[18] Daniel W Tigard. 2021. There is no techno-responsibility gap. *Philosophy & Technology* 34, 3 (2021), 589–607.

[19] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.