"decisionTime"
A Configurable Framework for Reproducible Human-AI Decision-Making Studies

Salimzadeh, Sara; Gadiraju, Ujwal

DOI
10.1145/3631700.3664885

Publication date
2024

Document Version
Final published version

Published in
UMAP 2024 - Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization

“DecisionTime”: A Configurable Framework for Reproducible Human-AI Decision-Making Studies

Sara Salimzadeh
Delft University of Technology
Delft, The Netherlands
s.salimzadeh@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@tudelft.nl

ABSTRACT

Empirical studies have extensively investigated human decision-making processes in various domains where AI systems are incorporated. However, comparing and replicating these studies can be challenging due to different experimental configurations. Moreover, the existing contexts often have limited scope and may not fully capture the complexity of real-world decision-making scenarios that are riddled with varying levels of uncertainty. Our framework addresses these practical gaps by providing a configurable and reproducible environment for conducting human-AI decision-making studies in the route planning domain that captures many complexities of real-world scenarios. Researchers can customize parameters, conditions, and factors involved in decision-making tasks to help address research and empirical gaps through rigorous experiments. With various modules such as map generation, chat components, and different AI systems available within the “`DecisionTime`” framework, researchers can effortlessly design experiments exploring multiple aspects of human-AI interaction and decision-making.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

ACM Reference Format:
Sara Salimzadeh and Ujwal Gadiraju. 2024. “DecisionTime”: A Configurable Framework for Reproducible Human-AI Decision-Making Studies. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct ’24), July 01–04, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3631700.3664885

1 INTRODUCTION

AI systems have been increasingly adapted in various domains to assist individuals in making decisions that directly impact their lives, such as allocating resources [4], recommending products [10], and even predicting medical conditions [12]. AI systems can potentially enhance the effectiveness of decision-making processes and the overall quality of outcomes as decision-support tools [3]. With AI systems having the potential to make significant real-world decisions that can significantly affect individual lives, it is essential to guarantee these systems’ trustworthiness, fairness, and accountability [11]. Researchers have investigated various factors affecting the interaction between humans and AI in decision-making, aiming to understand their impact on human behavior [7]. This understanding can contribute to developing approaches to enhance the human-AI decision-making processes.

Numerous empirical studies have explored the factors that influence interactions between humans and AI systems in decision-making situations, such as AI systems’ explainability [17], the accuracy of AI suggestions [8], the level of trust and reliance on AI systems [1, 9], and the decision context in which AI systems are employed [14]. Empirical studies involve human participants engaging in decision-making activities with assistance from AI systems. This allows researchers to observe and analyze their behavior, decision-making processes, and perceptions. Previous studies have also developed strategies, methods, and techniques to address existing challenges and utilize empirical investigations to evaluate the efficiency of these strategies. These methods may involve creating explainable AI systems [13], offering visualizations and interfaces that improve transparency and trustworthiness [5], and implementing evaluation metrics [16] to assess the impact of AI systems in decision-making processes.

While these empirical studies provide valuable insights into the dynamics of human-AI decision-making, there is a need for a more systematic and reproducible approach to designing experiments in this field [6]. This is crucial for laying a solid groundwork for advancing and assessing AI systems in decision-making. The context and domain of the decision-making tasks is a critical factor in designing empirical studies, as it determines the relevance and applicability of the findings to real-world scenarios [11]. It also influences the selection of appropriate evaluation metrics and experimental protocols [11]. For instance, the trustworthiness of AI systems would be evaluated differently in healthcare decision-making with high-stakes and potential harm compared to a low-stakes scenario like recommending movies. However, current decision tasks vary widely and lack standardization, making comparing results across studies challenging. T hey are often too simplified or artificially created, lacking ecological validity and real-world complexity [15].

To address these challenges, we propose “`DecisionTime`”, a configurable framework for designing reproducible studies in a route-planning context. Researchers and practitioners can manipulate and measure various variables in a controlled manner or extend the framework to incorporate more context-specific AI-related factors as needed. With these configurations, they can systematically investigate the impact of different variables on human-AI decision-making, including AI performance, system explainability, human

UMAP Adjunct ’24, July 01–04, 2024, Cagliari, Italy

ACM ISBN 979-8-4007-0466-6/24/07
https://doi.org/10.1145/3631700.3664885

trust and reliance, as well as contextual factors such as task complexity, time pressure, stakes, and information availability. Furthermore, a diverse range of participants with different levels of expertise can be engaged as individuals or in groups to perform tasks in route-planning domain. `DecisionTime` allows for a more precise representation of decision-making scenarios in terms of real-world complexity and user diversity, thereby enhancing the generalizability of the findings. Designing such a framework can be adapted for various decision-making domains to ensure reproducibility and enable meaningful comparisons across different studies. To the best of our knowledge, this is the first framework that addresses the need for systematic and reproducible experimental studies in human-AI decision-making using a configurable route-planning context.

The contribution of this work can be summarized as follows:

A **configurable** framework that enables researchers to design reproducible studies in route-planning contexts and potentially other decision-making domains. The framework will facilitate the design of experiments that closely align with real-world decision-making scenarios, increasing the ecological validity and generalizability of the findings.

A **modular** framework with adaptable components that can be customized to fit specific research questions and contextual factors. Our framework provides extensive control over the variables, allowing for systematic manipulation and comparison of various configurations. In prior research, experiments mainly have been conducted using the same dataset to train AI systems, resulting in limited flexibility.

2 ARCHITECTURE

Decision Domain: We selected route-planning as the specific decision domain for this framework due to several factors. First, route-planning is a common decision-making context in everyday life, making it relevant and relatable to many individuals. Participants of empirical studies can also perform route-planning tasks individually or collaboratively in groups, allowing for diverse participation and capturing of different decision-making dynamics. Second, route-planning involves considering multiple factors, such as time constraints and various possible route options, and evaluating cost, efficiency, and potential risks associated with routes. Various goals, such as minimizing travel time, avoiding toll roads, or optimizing fuel consumption, can be incorporated into the decision-making process, enabling a rich and diverse set of decision scenarios and tasks to be explored within this domain. Third, route-planning often involves the use of AI algorithms and technologies, making it a suitable domain for studying human-AI decision-making interactions and evaluating the impact of varying attributes in AI systems on decision outcomes.

`DecisionTime` framework comprises three central layers: a map and metadata generator, a web server, and a user interface. Researchers and practitioners can choose to utilize the layers separately or together to construct a tailored framework based on their specific research objectives and needs. **All code for this framework** is made publicly available to support future research in the community. [1]

Map and Metadata Generator: This is the core layer of the framework that generates realistic maps and relevant metadata that represent the decision-making task. The map and metadata generator utilizes spatial data and tailored functions to create maps with various features such as roads, landmarks, and points of interest. This metadata is based on OpenStreetMap data obtained using OSMnx [2], a Python library for retrieving, modeling, and analyzing street networks. The traffic, weather, and other related variables can also be incorporated into the generated maps to provide a more realistic environment for decision-making. Maps can be generated into HTML files, allowing easy integration with the web server layer, shown in Figure 1. The image of the generated map can also be exported for visualization purposes. Metadata is stored in JSON format, allowing researchers to access and manipulate the data for analysis and experimentation.

Web Server: The web server layer acts as the backbone of the framework, providing the infrastructure to support data storage, processing, and communication between different components. It handles requests from user interface, manages map data and metadata storage and retrieval, and facilitates communication with the user interface layer through the REST APIs. The existing APIs are programmed for decision-making studies, while they can be extended to include new functionalities based on specific research needs.

Interface: The user interface layer is responsible for providing a user-friendly interface through which participants of empirical studies can interact with the framework. It includes map visualization, route-planning, and decision-making functionalities. The design of the user interfaces enables researchers to use them as they are or customize them according to their specific study requirements. Additionally, the user interface allows researchers to collect data on participants' decision-making processes and interactions with the existing components.

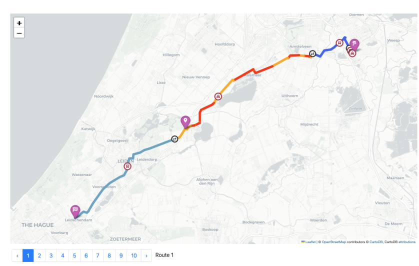

Figure 1: An illustration of the map within the interface, allowing participants to engage with the system and make decisions.

3 CUSTOMIZATION

Customization is a crucial aspect of the `DecisionTime` framework, allowing researchers and practitioners to tailor it to their needs and

[1] https://anonymous.4open.science/r/rp-364E/README.md

objectives. While every architectural layer provides default functionality and customization, the main contribution of this framework lies in the ability to customize the map and metadata generator, as well as the extensible user interface to widen the scope of decision-making studies.

3.1 Basic Map Configuration

The framework consists of several basic configurations for **map generation** that can be customized based on specific research requirements to ensure reproducibility and flexibility. These configurations include:

3.1.1 ***Route-specific attributes.*** : Researchers can identify the particular characteristics of a route that aim to be considered in the decision-making processes. These attributes may include distance, time, traffic conditions, road quality, and potential hazards. The choice of particular attributes can depend on the research questions and design choices. The presentation of attributes can also be tailored, for example, by displaying them as numerical values or graphical representations on a map. The initial and final points of the route, as well as any stops in between, can be adjusted to align with the researchers' needs. Additionally, customizing visual elements such as colors and icons for presenting routes or locations is essential for creating an intuitive and user-friendly interface for participants of empirical studies.

3.1.2 ***Map display.*** : `DecisionTime` framework provides different options for displaying the map interface for researchers. These options include different map styles (e.g., satellite view, street view) and zoom levels. Furthermore, researchers can include additional information on the map, such as points of interest, landmarks, and different types of roads (e.g., highways, local roads). The map display can also be interactive, allowing participants to zoom in and out and toggle between different layers of information. Maps can be formatted as HTML files to be displayed on web browsers or as images embedded in user interfaces or presentation materials.

3.2 Modular Map and Metadata Generator

Researchers are able to study how various factors affect human-AI decision-making by adapting the framework's map and metadata generator module. These factors can be divided into contextual elements and attributes of AI systems, which can be systematically manipulated to create different experimental scenarios. Contextual factors include the complexity of the decision-making task, the level of uncertainty in the environment, the presence of time constraints, risk factors, the availability of information, and the cognitive effort required to process the data. Additionally, attributes of the AI systems can be manipulated, such as the reliability of the AI predictions, the level of transparency in the decision-making process, the degree of autonomy given to the AI system, updating the knowledge of AI systems over time, and the level of user control or interaction with the AI system. By systematically manipulating and operationalizing these factors on the generated maps and metadata, researchers can study how different combinations of contextual elements and AI attributes influence human-AI decision-making outcomes.

3.3 In-Built Scaffolding

The framework contains in-built scaffolding modules that enable researchers to assist participants before the decision-making phase. These supportive modules are designed to minimize potential biases or misunderstandings, ensuring participants make decisions based on accurate information and a comprehensive understanding of the task. This ultimately leads to more reliable and valid data for analysis. The scaffolding components consist of tutorials, detailed guides, and hand-in experiences designed to assist participants in understanding the contextual factors and getting acquainted with the interface and navigation features. Standardizing these modules while adapting them to specific task details ensures consistency across different studies while maintaining consistent support for all participants. Additionally, we incorporate quiz questions as part of our sample assessment process to gauge participant comprehension of material covered in the scaffolding modules and their readiness for engaging in the decision-making process. These modules can be augmented, removed, or tailored based on the specific needs and requirements of the study, allowing for customization and flexibility in the implementation process.

3.4 Chat Component Extension

The user interface of `DecisionTime` supports a chat component extension that facilitates collaborative decision-making and group discussions. This feature enables groups of various sizes to engage in real-time conversations, exchange perspectives, and collectively reach consensus. Additionally, participants can use the chat to ask questions, seek clarifications, and receive timely feedback from facilitators or researchers. It also provides an avenue for sharing feedback on AI predictions or suggestions, contributing to a more interactive and dynamic decision-making process.

3.5 Online AI System Extension

Our framework can integrate multiple AI systems to suggest and generate decision-making options for participants according to their preferences, task objectives, and contextual information in offline and online settings. In offline scenarios, the AI system provides precomputed suggestions to participants during the decision-making process. This represents the default operation mode, during which maps and routes are pre-generated to facilitate efficient navigation and exploration. For real-time interactions in online setting, AI suggestions can be augmented into the user interface with live data feeds, allowing participants to receive up-to-date information and recommendations where maps and routes are dynamically generated based on real-time data.

4 CONCLUSION

In this work, we introduce our configurable framework, `DecisionTime`, for reproducible human-AI decision-making studies, which incorporates various modules to support effective decision-making processes. Our framework empowers researchers to design empirical studies tailored to their needs and research questions, highlighting the need for a systematic and rigorous approach to studying human-AI decision-making. It offers different layers of functionality and adaptability, from map generation to setting up the entire study flow. With scaffolding modules, a chat component extension, and an online AI system extension included, our framework provides a comprehensive tool-set for conducting human-AI decision-making studies. In the future, our goal is to improve our framework by broadening the scope of study scenarios and domain applications, enhancing data analysis capabilities, and incorporating additional templates for simple reproduction and adaptation for a larger community.

REFERENCES

[1] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (apr 2023), 17 pages. https://doi.org/10.1145/3579460

[2] Geoff Boeing. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, environment and urban systems* 65 (2017), 126–139.

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[4] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>College Station</city>, <state>TX</state>, <country>USA</country>, </conf-loc>) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[5] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, and Qian Pan. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 392–401. https://doi.org/10.1145/3397481.3450698

[6] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[7] Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 537–548. https://doi.org/10.1145/3490099.3511111

[8] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*.

[9] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. https://doi.org/10.1145/3544548.3581025

[10] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[11] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[12] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Interactive Hybrid Approach to Combine Machine and Human Intelligence for Personalized Rehabilitation Assessment. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (Toronto, Ontario, Canada) *(CHIL '20)*. Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/3368555.3384452

[13] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 97–105. https://doi.org/10.1609/hcomp.v7i1.5284

[14] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[15] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 215–227. https://doi.org/10.1145/3565472.3592959

[16] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).

[17] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605