A Frictional Design Approach: Towards Judicial AI and its Possible Applications

Caterina Fregosi[1,*], Federico Cabitza[1,2]

[1]Università degli Studi di Milano-Bicocca, Milan, Italy
[2]IRCCS Ospedale Galeazzi-Sant'Ambrogio, Milan, Italy

Abstract

Decision support systems (DSS) are increasingly being integrated into high-stakes domains like healthcare, law, and finance, where critical decisions have significant consequences. Traditional DSS often provide a single, clear-cut recommendation, which can lead to automation bias and diminish the user's sense of agency. However, there is a growing concern about the over-reliance on these systems and the potential for deskilling among users. The knowledge gap we aim to address is the development of decision support systems that effectively encourage critical reflection and maintain user engagement and responsibility in decision-making processes. In this workshop contribution, we report on the development of Judicial AI, a novel approach inspired by Frictional AI. Judicial AI diverges from traditional DSS by offering multiple, contrasting explanations to support different potential outcomes. This design encourages users to engage in deeper cognitive processing, thereby promoting critical reflection, reducing automation bias, and preserving the user's sense of agency. This ongoing study employs a two-arm experiment to investigate the effects of this approach in the context of content classification tasks, comparing it with the traditional protocol. The expected outcomes of this ongoing study suggest that the Judicial protocol could not only mitigate automation bias but also safeguard users' sense of agency and promote long-term skill retention.

Keywords

Frictional AI, Judicial AI, Human-AI Decision making process, eXplainable AI (XAI)

1. Introduction

In domains where decision-makers face high-stakes scenarios with significant consequences, Decision Support Systems (DSS) are increasingly implemented. It is essential to support users not only in identifying the optimal decision but also in effectively managing the decision-making process. This approach aims to mitigate the detrimental effects of interaction, such as overconfidence or underconfidence [1], while fostering appropriate reliance on the decision support system [2]. This involves providing users with support to critically assess both their own reasoning processes and the AI system's recommendations, a feature often absent in oracular decision support systems [3]. Such systems tend to offer clear-cut answers, thereby fostering an uncritical reliance on the system. Cooper (1999) introduced the concept of cognitive friction defined as "the resistance encountered by a human intellect when it engages with a complex

HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 10—14, 2024, Malmö, Sweden

*Corresponding author.
c.fregosi@campus.unimib.it (C. Fregosi); federico.cabitza@unimib.it (F. Cabitza)
0009-0004-7626-8131 (C. Fregosi); 0000-0002-4065-3415 (F. Cabitza)

system of rules" [4]. In technology domain, a design inspired by friction concept intentionally incorporates what Cox et al (2016) describe as design frictions "points of difficulty encountered during interaction with technology" [5] or what Cabitza et al (2019) term programmed inefficiencies [6]. Contrary to trends aiming to create seamless interactions that promote speed and efficiency, a "positive friction "strategy deliberately integrates these elements to improve user engagement and reflection [7]. The term Frictional AI was introduced by Cabitza et al (2024) as an umbrella term for a variety of methods aimed at encouraging reflection in human-AI decision making processes by introducing cognitive friction [8].

In the domain of Human-Computer Interaction (HCI), the design of decision support systems (DSS) that offer multiple, well-argued explanations for different hypotheses—rather than simply presenting the (allegedely) correct answer—presents two significant advantages that address cognitive and ethical concerns. Firstly, such a system is designed to mitigate the risk of automation bias, a well-documented phenomenon where users over-rely on automated systems [9], often accepting their outputs uncritically even when they are wrong [10]. By presenting multiple plausible explanations, backing up each option, the DSS compels the user to engage in deeper cognitive processing, comparing and contrasting the arguments put forth for each hypothesis. This engagement naturally limits automation bias, as users are less likely to defer uncritically to a single system-provided solution. Even when one explanation appears more convincing, the presence of alternative perspectives serves as a safeguard, ensuring that the user remains critical, reducing (but not eradicating) the chances of endorsing a false or irrelevant conclusion.
Secondly, offering multiple explanations helps address a less explored but equally important issue in HCI: the potential loss of agency in human-AI interaction [11, 12], especially when the system is renown for its accuracy and reliability. When users are presented with only one "right" answer, they may gradually lose their sense of control and responsibility over decision-making processes [13]. This phenomenon, which can be assimilated to the concept of deresponsibilization [14], reflects the risk that users may start to perceive themselves as mere executors of the system's decisions rather than active, responsible agents, which are accountable for the final decision (as they still are). Over time, this can lead to long-term consequences, including loss of motivation, loss of skill and hampered learning [8]. By fostering an environment where the user must evaluate and decide between multiple, well-supported hypotheses, the DSS preserves and even enhances the user's sense of agency. The user remains an active participant in the decision-making process, fully responsible for the final choice, which in turn helps maintain and develop their cognitive skills.

To this end, we have designed an experiment introducing a Judicial system, one of the protocol associated with Frictional AI, which involves an AI system providing contrasting plausible explanations that each support a different decision outcome [15, 8].

This resonates with the "agonistic machine learning" models [16] and with the Evaluative AI paradigm introduced by Miller (2023) [3] for explainable decision support. The novelty introduced by the Judicial AI system is that, inspired by the judicial domain, it proposes distinct explanations to support each of the two possible outcomes.

In this project we investigate the textual generative setting in Judicial protocol for sentence classification and its effects in terms of accuracy, confidence, reliance, perceived responsibility

and sense of agency for the decisions made.

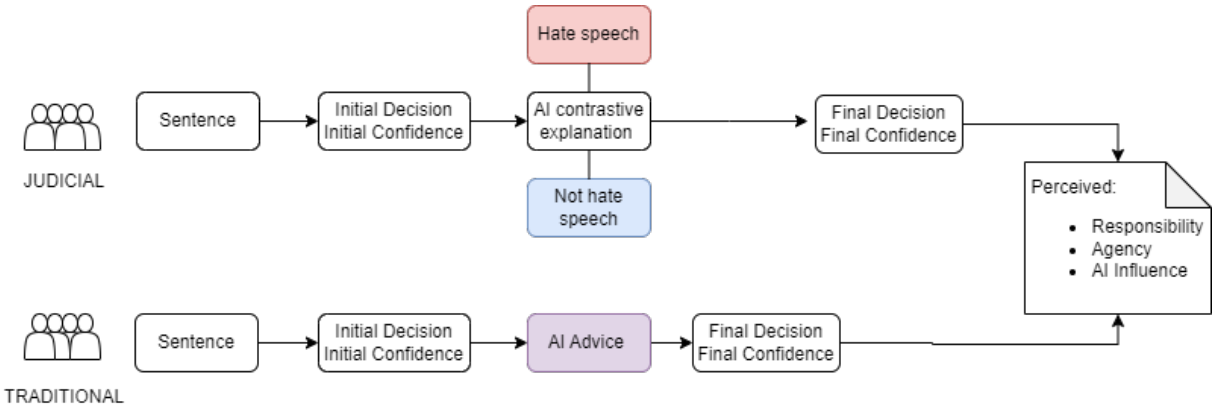

Figure 1: A schematic view of the study design. Participants are divided into two groups: one group receives AI contrastive explanations (*Judicial Protocol*), while the other group receives AI advice (*Traditional Protocol*). Both groups make an initial decision and indicate their initial confidence level before interacting with the AI. After receiving the AI input, participants make a final decision and report their final confidence. The study also assesses participants' perceived responsibility, sense of agency, and AI influence.

2. Methods

As illustrated in 1, we will conduct a two-arm experiment to examine how different interaction designs influence decision-making in content moderation tasks. Participants in each arm will be presented with the same set of 30 sentences, previously identified as complex by state-of-the-art hate speech detection systems and sourced from a social media platform. The two arms of the study will employ distinct interaction protocols, which will be randomly assigned to participants.

- **Judicial**: Participants will be presented with a sentence and asked to classify it as either hate speech or not hate speech. Additionally, they will be required to rate their confidence in their decision using a four-level ordinal scale, which ranges from "not at all confident" to "completely confident". Following this initial judgment, participants will be shown the sentence again, accompanied by two arguments, generated by the Judicial AI system, presented in colored boxes: one in a pastel red box on the left, advocating for the classification of the sentence as hate speech, and another in a pastel blue box on the right, presenting an opposite argument. Participants will then be asked to provide their final decision and confidence rating using the same scale as before. This process will be systematically repeated for all 30 cases. To minimize potential order bias in the decision-making process, half of the participants assigned this protocol will encounter first the opposing viewpoint. Specifically, the pastel blue boxes, representing arguments for content classification as not hate speech, will be positioned on the left, while the pastel red boxes, supporting the classification of content as hate speech, will be on the right.

- **Traditional**: The initial screen for each case will be consistent with the Judicial arm. Participants will first be asked to make an initial decision on whether the sentence constitutes hate speech or not, and to rate their confidence in this decision. After this, the system will present its classification of the sentence (hate speech or not) and participants will be asked to either confirm or reject this classification, providing their confidence level in this final decision.

Pre-test and post-test questionnaires will be administered to assess participants' trust in the AI system. Additionally, the post-test questionnaire will evaluate the sense of agency and responsibility participants perceive regarding the decisions they made during the study.

We expect to address the following research questions:
R1: Is the Traditional protocol associated with higher accuracy compared to the Judicial protocol?
R2: Are respondents of the Judicial protocol more confident than Traditional ones in their own final decision?
R3: Is there a significant difference in reliance behavior between the Judicial and Traditional protocols?
R4: Do Judicial respondents feel a higher sense of agency and responsibility regarding their decisions compared to Traditional respondents?

To address the proposed research questions, a series of analyses will be conducted [1] on the groups of users subjected to the Traditional and Judicial interaction protocols, as outlined in Table 1.

3. Expected results

A DSS that offers multiple plausible explanations not only aligns with the principles of user-centered design but also plays a crucial role in maintaining critical engagement, preserving user agency, and ensuring the retention of decision-making skills, thereby addressing both automation bias and the risk of deskilling in human-AI interactions. By encouraging deeper and more critical reflection, this design reduces the risk of fostering undue user trust, which can contribute to the White Box Paradox [17]. However, it is important to note that the protocol could still inadvertently introduce bias if one explanation seems more convincing, even if it is incorrect. The adoption of the Judicial protocol in human-AI interaction is expected to have a significant impact on the quality of decisions made by users, in particular on perceived agency and control over their choices. Therefore, we believe Judicial AI could represent a promising direction in the study of improved decision support system processes, potentially increasing both the effectiveness of these systems and user satisfaction. Further research focused on refining Judicial protocols and examining their long-term effects could have significant implications for the design and implementation of future decision support systems.

[1] with the tool available at https://mudilab.github.io/dss-quality-assessment/

Table 1
Research questions (RQs) and corresponding planned analyses to evaluate the impact of Judicial and Traditional protocols on user accuracy, confidence, reliance, and sense of agency in decision-making processes.

RQ	Research Question	Planned analysis
RQ1	Is the Traditional protocol associated with higher accuracy compared to the Judicial protocol?	Compare initial and final accuracy levels of users in both the Traditional and Judicial protocol groups to assess if the absence of direct advice in the Judicial protocol affects final accuracy.
RQ2	Are respondents of the Judicial protocol more confident than Traditional ones in their own final decision?	Analyze and compare the final confidence levels and the differences between initial and final confidence for both groups to determine if the Judicial protocol increases confidence in the final decision.
RQ3	Is there a significant difference in reliance between the Judicial and Traditional protocols?	Compare reliance by analyzing how users in both groups rely on correct or incorrect advice/explanations. Examine cases where the initial decision differs from the final decision to identify reliance patterns.
RQ4	Do Judicial respondents feel a higher sense of agency and responsibility regarding their decisions compared to Traditional respondents?	Analyze the final questionnaire responses to compare the perceived levels of AI influence, responsibility, and sense of agency between the two groups.

Acknowledgments

C. Fregosi and F. Cabitza acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

References

[1] T. Kliegr, Š. Bahník, J. Fürnkranz, A review of possible effects of cognitive biases on interpretation of rule-based machine learning models, Artificial Intelligence 295 (2021) 103458.

[2] F. Cabitza, A. Campagner, R. Angius, C. Natali, C. Reverberi, Ai shall have no dominion: on how to measure technology dominance in ai-supported human decision-making, in: Proceedings of the 2023 CHI conference on human factors in computing systems, 2023, pp. 1–20.

[3] T. Miller, Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 333–342.

[4] A. Cooper, The inmates are running the asylum, Springer, 1999.

[5] A. L. Cox, S. J. Gould, M. E. Cecchinato, I. Iacovides, I. Renfree, Design frictions for mindful

interactions: The case for microboundaries, in: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems, 2016, pp. 1389–1397.
[6] F. Cabitza, A. Campagner, D. Ciucci, A. Seveso, Programmed inefficiencies in dss-supported human decision making, in: Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, September 4–6, 2019, Proceedings 16, Springer, 2019, pp. 201–212.
[7] Z. Chen, R. Schmidt, Exploring a behavioral model of" positive friction" in human-ai interaction, arXiv preprint arXiv:2402.09683 (2024).
[8] F. Cabitza, C. Natali, L. Famiglini, A. Campagner, V. Caccavella, E. Gallazzi, Never tell me the odds: Investigating pro-hoc explanations in medical decision making, Artificial Intelligence in Medicine (2024) 102819.
[9] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–21.
[10] M. Vered, T. Livni, P. D. L. Howe, T. Miller, L. Sonenberg, The effects of explanations on automation bias, Artificial Intelligence 322 (2023) 103952.
[11] H. Limerick, J. W. Moore, D. Coyle, Empirical evidence for a diminished sense of agency in speech interfaces, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 3967–3970.
[12] A. Galsgaard, T. Doorschodt, A.-L. Holten, F. C. Müller, M. P. Boesen, M. Maas, Artificial intelligence and multidisciplinary team meetings; a communication challenge for radiologists' sense of agency and position as spider in a web?, European Journal of Radiology 155 (2022) 110231.
[13] R. Legaspi, W. Xu, T. Konishi, S. Wada, N. Kobayashi, Y. Naruse, Y. Ishikawa, The sense of agency in human–ai interactions, Knowledge-Based Systems 286 (2024) 111298.
[14] C. Sureau, Medical deresponsibilization, Journal of assisted reproduction and genetics 12 (1995) 552–558.
[15] C. Natali, et al., Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems, in: CEUR workshop proceedings, volume 3481, 2023, pp. 15–19.
[16] M. Hildebrandt, Privacy as protection of the incomputable self: From agnostic to agonistic machine learning, Theoretical Inquiries in Law 20 (2019) 83–121.
[17] F. Cabitza, A. Campagner, L. Ronzio, M. Cameli, G. E. Mandoli, M. C. Pastore, L. M. Sconfienza, D. Folgado, M. Barandas, H. Gamboa, Rams, hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine 138 (2023) 102506.