Frictional AI in Joint Cognitive Systems: Towards a Human-Centered Approach at Higher Levels

Andrea La Rosa[1,*,†], Andrea Beretta[2,†]

[1]University of Padua, 35131 Padova, Italy
[2]Institute of Information Science and Technologies (ISTI-CNR), 56124 Pisa, Italy

Abstract
Artificial intelligence systems are increasingly embedded in everyday and high-stakes decision making. However, the pace and seamlessness of human-AI interactions can undermine critical reflection and meaningful human control. This concern becomes especially relevant in complex sociotechnical systems, where multiple human and machine actors must coordinate across interdependent levels to achieve shared goals. Drawing on a human-centered design perspective, we frame this cooperation through the lens of the Joint Cognitive System (JCS) theory, which conceptualizes all actors — human and artificial — as components of a unified cognitive entity. Within this context, we examine the potential integration of friction mechanisms into complex systems, i.e., intentional design constraints that slow down or challenge interaction to promote deliberation, human control, and appropriate trust and reliance. We argue that building friction in interaction can improve decision quality while preserving human oversight at every level of the sociotechnical system - as long as properly adapted and scaled according to the its functional role and degree of control. We discuss theoretical foundations, outline design guidelines, and identify research directions to effectively address challenges and opportunities arising from AI-driven technological advancements.

Keywords
Human-Centered Explainable AI, Sociotechnical Systems, Joint Cognitive Systems, Human–AI Decision Making, Frictional AI, Human–AI Interaction Design

1. Introduction

Artificial intelligence (AI) systems are rapidly becoming woven into both routine and high-stakes decision processes (e.g., educational, financial and healthcare domains, but also disaster management or diverse range of Generative AI applications [1][2][3][4][5]) changing not only the speed and scale of decisions but also how responsibility and cognitive work are distributed across people and machines. At the same time, the pace, automation, and often seamless nature of many AI-mediated interactions can reduce opportunities for deliberation, encourage superficial acceptance (or even total skepticism) of machine suggestions, and erode meaningful human control [6][7].

However, as emphasized by human-centered perspectives, these systems should enhance or augment human capabilities *while* preserving human judgment, oversight, and agency [8][9][10]. Consequently, the concept of Explainable Artificial Intelligence (XAI) has been the main focus in optimizing human-AI interactions, aiming to enhance system transparency and reliability while promoting a human-centered design approach (HCXAI) [11][12][13].

While XAI would provide interpretability for *individual* systems, these dynamics are particularly significant in **complex sociotechnical systems**, where multiple human and artificial actors must coordinate across levels of organization and uncertainty [14][15][16]. Within this context, we argue that the **Joint Cognitive System (JCS)** theory represents the key framework to observe how human and AI co-evolve [17][18]. In JCS cognition is not confined to a single agent (whether human or machine), but

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy
[*]Corresponding author.
[†]These authors contributed equally.
andrea.larosa.1@studenti.unipd.it (A. La Rosa); andrea.beretta@isti.cnr.it (A. Beretta)
0000-0001-8531-9325 (A. Beretta)

rather emerges from the collaboration between multiple actors operating within a specific context and across different level of analysis (e.g., how humans and AI systems interact in dynamic and complex decision-making scenarios).

In parallel, the suggestion of leveraging **friction in AI** has been proposed [6][19] as a design apparatus to introduce deliberate constraints to promote user awareness and control, departing from the most used model of a fully seamless but potentially bias-inducing interface [20]. Friction, in this context, refers not to usability barriers but to cognitive and decision-making resistance that encourages reflection, agency, improves shared understanding, and helps mitigate overreliance on automation [6][20][21].

Although a promising approach, the existing literature has predominantly focused on designing constraints for a simplified, 'closed' *human-AI dyad* (a single human interacting with a single AI system), operating in a narrowly defined and controlled environment [6][21][22].

This raises a critical question about the effectiveness of such mechanisms as systems evolve into a higher-order JCS: can friction models developed for a single human-AI dyad be straightforwardly extended to complex sociotechnical decision-making environments (see Figure 1), or is it necessary to develop new techniques that account for the inherent complexity of multi-level JCS?

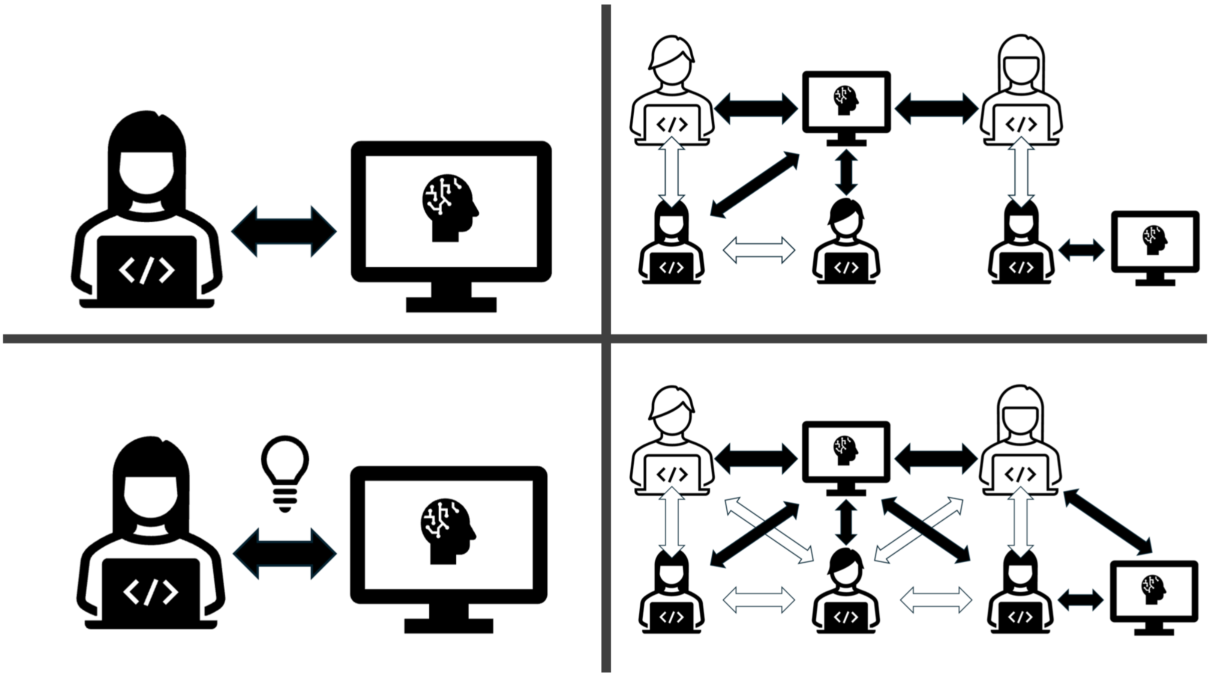

Figure 1: Joint Cognitive Systems and Friction in Decision Making: **top-left** depicts a simple, 'closed' human-AI dyad; **bottom-left** represents the core idea behind JCS, i.e., cognition emerging from interaction; **on the right**, two *different* sociotechnical *structures* with the *same number of actors* involved - a deliberate choice to further underline the hidden complexity and variety of possible interactions. **Black arrows** symbolize friction mechanisms, whereas **white arrows** are seamless connections between human actors: the guiding principle is to enhance human capabilities, while keeping human agency and without affecting system agility.

2. Human-AI Decision Making and Frictional AI

Within the field of intelligent decision support system (IDSS), i.e., AI-powered decision support system (DSS)[23][24], two concepts are widely regarded as fundamental: human-in-the-loop and human-in-control [25][26][27][28]. These paradigms closely align with the principles acknowledged and adopted by HCXAI researchers - as both aim to ensure meaningful human control throughout the whole AI system's lifecycle, from model training and deployment to ongoing use and iterative refinement.

Our work specifically addresses the *'choose among alternatives'* phase in decision-making processes

[29][30] by introducing *friction by design* as a deliberate strategy to preserve human agency and control in IDSSs, while mitigating the emergence of inappropriate reliance [20]. Through a combination of theoretical analysis and experimental testing, primarily conducted in controlled one-on-one environments, several friction-based techniques and protocols have been identified [22][6] - even though we note that the *ecological validity* of these empirical findings may be constrained by the experimental context.

Following [22] classification, they differentiate between:

- *Cautious protocols*, where the IDDS can either presents multiple choices, each associated with a 'confidence score', or none ('abstention') if the problem is too complex (to avoid deceiving the user, hence 'cautious');
- *Judicial or antagonist protocols*, where one or multiple IDSSs sustain and defend different or even opposite options - in order to promote human deliberation;
- *Decentralized AI or adjunct protocols*, where friction is used to encourage free, autonomous thinking *before* interacting with the IDSS (e.g., Cognitive Forcing Functions by [6]: having to wait n seconds before getting an answer, AI system acting as second-opinion giver solely, or having to explicitly request AI assistance);
- *Comparative or analogical protocols*, where the IDSS highlights the most analogous cases for each suggested alternative, providing the user with a more robust *context* for making informed choices.

Notably, each protocol entails distinct advantages and limitations, as well as suitable areas of application. For example, in time-critical contexts, adjunct techniques may prove ineffective or even counterproductive (e.g., in a clinical setting [31]), whereas a comparative approach might help expert clinicians. The cautious protocol, on the other hand, is susceptible to estimation errors that may exacerbate AI misuse and underuse [32], although the useful 'abstention' feature. Finally, antagonist protocol could significantly promote user awareness and healthy habits, as shown by a recent study [33].

3. A Design Perspective on Friction in Higher-Level JCS

When scaled to broader sociotechnical systems (e.g., healthcare, aviation, policy-making, disaster management; see Figure 2) [34], JCS analysis suggests that agency and control, as well as the related concept of 'responsibility', vary significantly across system layers, reflecting the escalating complexity and heterogeneity of actors and dynamics involved. What constitutes useful friction at a micro level (e.g., Cognitive Forcing Functions [6]) may be ineffective or even counterproductive at a macro level (e.g., multi-stakeholder decision-making in public policy). So, how should we address this gap without flattening complexity or oversimplifying human-AI dynamics?

To achieve this, our proposal encompasses several key design considerations:

- **Scalability of friction mechanisms**: in multi-level JCS, friction mechanisms must be designed hierarchically to regulate both inter- and intra-level interactions. A hierarchical model of friction enables context-sensitive modulation of its intensity and form, aligned with the decision-making authority at each level. This approach should in the end support transparency, accountability, and human oversight - all without compromising system agility;
- **Iterative and adaptive design guided by human-centered principles**: in complex dynamic decision-making scenerios, friction cannot be a static feature. Inspired by the *muddling through* concept [15], we suggest an iterative design approach in which friction mechanisms are continuously refined based on empirical evaluations (e.g., system performance metrics) and user feedback, establishing a *feedback loop* [35][36][37]. Such adaptive systems would dynamically modulate the intensity and type of friction applied, ensuring that the joint cognitive system remains both resilient and human-centered, while not overloading the cognitive capacity of human operators;
- **New needs, new variables, new metrics**: critically, existing studies on Frictional AI, such as [6][22], have primarily focused on tightly scoped, micro-level interactions. These findings,

while valuable, must be expanded through empirical and design research that explores friction in distributed, multi-level contexts. However, in order to properly do this, it is essential to identify which variables and metrics are relevant to these broader contexts. The challenge, therefore, is to design frictional protocols that are context-sensitive, adapting to the structure and scale of the JCS.

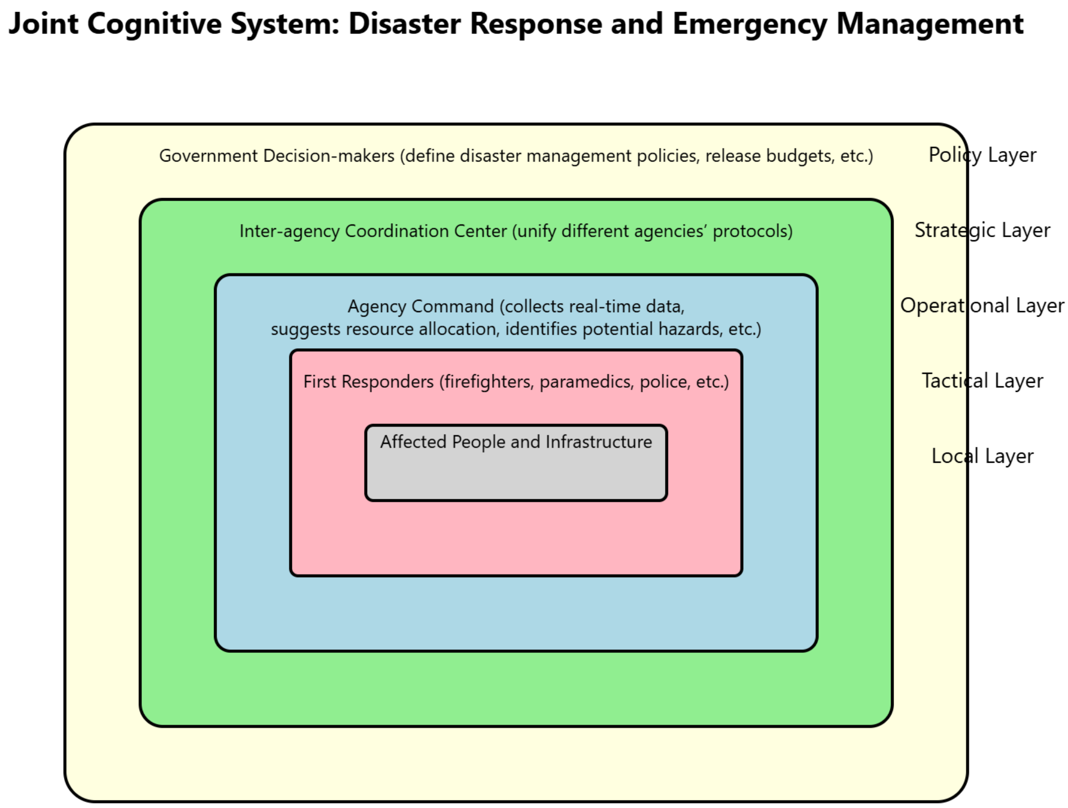

Figure 2: Example of a Joint Cognitive System for disaster response and emergency management. Notice the hierarchical structure of the different layers of intervention (with their sample name on the right): **each one is a JCS *per se***, with both human and technological actors, **as well as being part of a broader JCS**. In this example, each level is provided with a short description of people involved; one can however easily figure out AI-based systems that might help the human counterpart in various ways at any stage. For readers interested in a detailed discussion of this topic, see [4]. The present figure is shown only as an illustrative application of the framework developed here, and is not related to that line of research.

4. Conclusion and Future Directions

This work opens a research agenda on the purposeful design and use of friction mechanisms in AI-enabled multi-level Joint Cognitive Systems.

Given their complex nature, our hypothesis is that friction strategies effective in dyadic human-AI interactions do not directly generalize to multi-actor, multi-layer sociotechnical systems in which control and decision-making processes are distributed across varied roles and layers. Instead, we propose friction as a scalable, adaptable design element within a human-centered view of JCS to support more efficient and effective decision-making, keeping humans in-the-loop and in-control while promoting appropriate reliance.

As an initial step toward this agenda, we offer preliminary insights and design guidelines to support future research and practice. We also identify and recommend three directions for further investigation:

- Development of frameworks for identifying appropriate friction points at different levels of a JCS;
- Case studies that evaluate frictional design in multi-agent, multi-layered contexts;
- New evaluation metrics that go beyond user satisfaction to measure long-term impact on decision quality, system resilience, as well as human agency and coordination at different JCS scopes.

This work is subject to limitations that arise from its scope and underlying assumptions. First, we deliberately focus on providing theoretical considerations and guidelines according to solid design principles [15][17][20] rather than translating into more concrete design patterns. There are two main reasons behind this choice: a lack of real-world applications' data and the high variability inherent in a complex sociotechnical system (e.g., differences in structure, needs, communication strategies, degree of control).

Secondly, we choose to adopt the Joint Cognitive Systems view as our main theoretical framework. Although we advocate for this decision, conceptualizing AI components as cognitive 'teammates' [18][38] within a Joint Cognitive System could inadvertently encourage anthropomorphism [39][40], which in turn may inflate trust and foster inappropriate reliance.

Regarding friction, it should be noted that 'friction' itself may not be universally recognized as the standard term for this concept (e.g., 'Seamful' XAI [41]). Consequently, we underscore the need for a consistent, widely recognized terminology to support research. Furthermore, while friction mechanisms may prove to be beneficial, their application requires careful calibration. Critically, poorly designed or excessive interventions can induce frustration [6][33] and even habituation effects on the human side [42]. Accordingly, frictional mechanisms should be designed and evaluated under explicit human-oversight and appropriate-reliance objectives, ensuring that their deployment sustains human agency and system resilience while remaining in compliance with contemporary local regulations for AI systems (e.g., high-risk AI systems [43]).

Acknowledgments

This work has been funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI".

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] W. Forero-Corba, F. N. Bennasar, Techniques and applications of machine learning and artificial intelligence in education: a systematic review, RIED-Revista Iberoamericana de Educación a Distancia 27 (2024).
[2] P. Weber, K. V. Carl, O. Hinz, Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature, Management Review Quarterly 74 (2024) 867–907.
[3] H. A. Younis, T. A. E. Eisa, M. Nasser, T. M. Sahib, A. A. Noor, O. M. Alyasiri, S. Salisu, I. M. Hayder, H. A. Younis, A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges, Diagnostics 14 (2024) 109.
[4] A. Albahri, Y. L. Khaleel, M. A. Habeeb, R. D. Ismael, Q. A. Hameed, M. Deveci, R. Z. Homod, O. Albahri, A. Alamoodi, L. Alzubaidi, A systematic review of trustworthy artificial intelligence applications in natural disasters, Computers and Electrical Engineering 118 (2024) 109409.
[5] S. S. Sengar, A. B. Hasan, S. Kumar, F. Carroll, Generative artificial intelligence: a systematic review and applications, Multimedia Tools and Applications 84 (2025) 23661–23700.

[6] 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, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://doi.org/10.1145/3449287. doi:10.1145/3449287.

[7] O. Turel, S. Kalhan, Prejudiced against the machine? implicit associations and the transience of algorithm aversion., MIS quarterly 47 (2023).

[8] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504.

[9] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al., Guidelines for human-ai interaction, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–13.

[10] A. Hanif, X. Zhang, S. Wood, A survey on explainable artificial intelligence techniques and challenges, in: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), 2021, pp. 81–89. doi:10.1109/EDOCW52865.2021.00036.

[11] Z. Buçinca, A. Chouldechova, J. W. Vaughan, K. Z. Gajos, Beyond end predictions: stop putting machine learning first and design human-centered ai for decision support, in: HCAI@NeurIPS, 2022.

[12] U. Ehsan, P. Wintersberger, Q. V. Liao, M. Mara, M. Streit, S. Wachter, A. Riener, M. O. Riedl, Operationalizing human-centered perspectives in explainable ai, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411763.3441342. doi:10.1145/3411763.3441342.

[13] U. Ehsan, P. Wintersberger, Q. V. Liao, E. A. Watkins, C. Manger, H. Daumé III, A. Riener, M. O. Riedl, Human-centered explainable ai (hcxai): Beyond opening the black-box of ai, in: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491101.3503727. doi:10.1145/3491101.3503727.

[14] E. Trist, W. Pasmore, J. Sherwood, Socio-technical systems, London: Tavistock (1960).

[15] D. A. Norman, P. J. Stappers, Designx: Complex sociotechnical systems, She Ji: The Journal of Design, Economics, and Innovation 1 (2015) 83–106. URL: https://www.sciencedirect.com/science/article/pii/S240587261530037X. doi:https://doi.org/10.1016/j.sheji.2016.01.002.

[16] D. Polojärvi, E. Palmer, C. Dunford, A systematic literature review of sociotechnical systems in systems engineering, Systems Engineering 26 (2023) 482–504. URL: https://incose.onlinelibrary.wiley.com/doi/abs/10.1002/sys.21664. doi:https://doi.org/10.1002/sys.21664. arXiv:https://incose.onlinelibrary.wiley.com/doi/pdf/10.1002/sys.21664.

[17] E. Hollnagel, Designing for joint cognitive systems, in: IEE and MOD HFI DTC Symposium on People and Systems-Who are we Designing for?, IET, 2005, pp. 47–51.

[18] W. Xu, Z. Gao, Applying hcai in developing effective human-ai teaming: A perspective from human-ai joint cognitive systems, Interactions 31 (2024) 32–37.

[19] C. Natali, Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems, in: G. Varni, G. Volpe (Eds.), Proceedings of the Doctoral Consortium of the 15th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2023), Turin, Italy, September 21st, 2023, volume 3481 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 15–19. URL: https://ceur-ws.org/Vol-3481/paper3.pdf.

[20] 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, CHI EA '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1389–1397. URL: https://doi.org/10.1145/2851581.2892410. doi:10.1145/2851581.2892410.

[21] F. Cabitza, C. Fregosi, A. Campagner, C. Natali, Explanations considered harmful: The impact of misleading explanations on accuracy in hybrid human-ai decision making, in: L. Longo, S. Lapuschkin, C. Seifert (Eds.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2024, pp. 255–269.

[22] 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 150 (2024) 102819. URL: https://www.sciencedirect.com/science/article/pii/S0933365724000617. doi:https://doi.org/10.1016/j.artmed.2024.102819.
[23] G. Onwujekwe, H. R. Weistroffer, Intelligent decision support systems: An analysis of the literature and a framework for development, Information Systems Frontiers (2025) 1–32.
[24] B. Sahoh, A. Choksuriwong, The role of explainable artificial intelligence in high-stakes decision-making systems: a systematic review, Journal of Ambient Intelligence and Humanized Computing 14 (2023) 7827–7843.
[25] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artificial Intelligence Review 56 (2023) 3005–3054.
[26] R. Parasuraman, T. B. Sheridan, C. D. Wickens, A model for types and levels of human interaction with automation, IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans 30 (2000) 286–297.
[27] S. Niehaus, M. Hartwig, P. H. Rosen, S. Wischniewski, An occupational safety and health perspective on human in control and ai, Frontiers in artificial intelligence 5 (2022) 868382.
[28] B. Wagner, Liable, but not in control? ensuring meaningful human agency in automated decision-making systems, Policy & Internet 11 (2019) 104–122.
[29] H. A. Simon, A behavioral model of rational choice, The quarterly journal of economics (1955) 99–118.
[30] Y. R. Shrestha, S. M. Ben-Menahem, G. Von Krogh, Organizational decision-making structures in the age of artificial intelligence, California management review 61 (2019) 66–83.
[31] E. Rosbach, J. Ammeling, S. Krügel, A. Kießig, A. Fritz, J. Ganz, C. Puget, T. Donovan, A. Klang, M. C. Köller, P. Bolfa, M. Tecilla, D. Denk, M. Kiupel, G. Paraschou, M. K. Kok, A. F. H. Haake, R. R. de Krijger, A. F.-P. Sonnen, T. Kasantikul, G. M. Dorrestein, R. C. Smedley, N. Stathonikos, M. Uhl, C. A. Bertram, A. Riener, M. Aubreville, "when two wrongs don't make a right" - examining confirmation bias and the role of time pressure during human-ai collaboration in computational pathology, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3713319. doi:10.1145/3706598.3713319.
[32] J. Li, Y. Yang, R. Zhang, Y.-c. Lee, Overconfident and unconfident ai hinder human-ai collaboration, arXiv preprint arXiv:2402.07632 (2024).
[33] Z. Buçinca, S. Swaroop, A. E. Paluch, F. Doshi-Velez, K. Z. Gajos, Contrastive explanations that anticipate human misconceptions can improve human decision-making skills, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, 2025. URL: https://arxiv.org/abs/2410.04253.
[34] U. Ehsan, K. Saha, M. De Choudhury, M. O. Riedl, Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–32. URL: http://dx.doi.org/10.1145/3579467. doi:10.1145/3579467.
[35] D. Sjödin, V. Parida, M. Palmié, J. Wincent, How ai capabilities enable business model innovation: Scaling ai through co-evolutionary processes and feedback loops, Journal of Business Research 134 (2021) 574–587. URL: https://www.sciencedirect.com/science/article/pii/S0148296321003386. doi:https://doi.org/10.1016/j.jbusres.2021.05.009.
[36] M. Glickman, T. Sharot, How human–ai feedback loops alter human perceptual, emotional and social judgements, Nature Human Behaviour 9 (2025) 345–359.
[37] E. Katsamakas, O. Pavlov, Ai and business model innovation: Leveraging the ai feedback loop, Journal of Business Models (JOBM) 8 (2020) 22–30. URL: https://hdl.handle.net/10419/318931. doi:10.5278/ojs.jbm.v8i2.3532.
[38] R. Zhang, W. Duan, C. Flathmann, N. McNeese, G. Freeman, A. Williams, Investigating ai teammate communication strategies and their impact in human-ai teams for effective teamwork, Proc. ACM Hum.-Comput. Interact. 7 (2023). URL: https://doi.org/10.1145/3610072. doi:10.1145/3610072.

[39] M. Li, A. Suh, Anthropomorphism in ai-enabled technology: A literature review, Electronic Markets 32 (2022) 2245–2275.
[40] M. Blut, C. Wang, N. V. Wünderlich, C. Brock, Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai, Journal of the academy of marketing science 49 (2021) 632–658.
[41] U. Ehsan, Q. V. Liao, S. Passi, M. O. Riedl, H. Daumé, Seamful xai: Operationalizing seamful design in explainable ai, Proc. ACM Hum.-Comput. Interact. 8 (2024). URL: https://doi.org/10.1145/3637396. doi:10.1145/3637396.
[42] J. Bogon, S. Hößl, C. Wolff, N. Henze, D. Halbhuber, Cognitive integration of delays: Anticipated system delays slow down user actions, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3713475. doi:10.1145/3706598.3713475.
[43] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), PE/24/2024/REV/1, OJ L, 2024/1689, 12.7.2024, ???? Document 32024R1689, ELI: http://data.europa.eu/eli/reg/2024/1689/oj.