Never tell me the odds: Investigating pro-hoc explanations in medical decision making

Federico Cabitza[1,2,†], Chiara Natali[1], Lorenzo Famiglini[1], Andrea Campagner[2], Valerio Caccavella[3], and Enrico Gallazzi[3]

[1]Università degli Studi di Milano-Bicocca, Milan, Italy
[2] IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
[3] Istituto Ortopedico Gaetano Pini — ASST Pini-CTO, Milan, Italy
[†] Corresponding author: federico.cabitza@unimib.it

Abstract. This paper examines a kind of explainable AI, centered around what we term *pro-hoc explanations*, that is a form of support that consists of offering alternative explanations (one for each possible outcome) *instead of* a specific *post-hoc* explanation following specific advice. Specifically, our support mechanism utilizes *explanations by examples*, featuring analogous cases for each category in a binary setting. Pro-hoc explanations are an instance of what we called *frictional AI*, a general class of decision support aimed at achieving a useful compromise between the increase of decision effectiveness and the mitigation of cognitive risks, such as over-reliance, automation bias and deskilling. To illustrate an instance of frictional AI, we conducted an empirical user study to investigate its impact on the task of radiological detection of vertebral fractures in x-rays. Our study engaged 16 orthopedists in a 'human-first, second-opinion' interaction protocol. In this protocol, clinicians first made initial assessments of the x-rays without AI assistance and then provided their final diagnosis after considering the pro-hoc explanations. Our findings indicate that physicians, particularly those with less experience, perceived pro-hoc XAI support as significantly beneficial, even though it did not notably enhance their diagnostic accuracy. However, their increased confidence in final diagnoses suggests a positive overall impact. Given the promisingly high effect size observed, our results advocate for further research into pro-hoc explanations specifically, and into the broader concept of frictional AI.

Keywords: eXplainable AI, Decision Support, Machine Learning, Frictional AI

1 Introduction

One of the earliest and most influential works promoting the human-centered approach to the design of interactive computer systems is Norman's book 'The Design of Everyday Things' [56]. Based on various studies conducted in the early years of personal computing, Norman promoted the basic principles of a design philosophy that would consider the needs, preferences, and requirements of users to make their use experiences not only more effective but also enjoyable, and

consequently to ensure better efficacy of action through pleasantness and tool usability. From this idea, a very broad consensus emerged that the interfaces of digital tools should be developed to make the use of the systems as natural as possible, intuitive, easy, without barriers and difficulties: "Don't make me think" by Krug [46] soon became a bestseller in the community of user interface designers and Human-Computer Interaction (HCI) scholars.

Yet, at the same time, Norman also warned that systems should not become *too easy to use*, because this could make users complacent and unthinking in their interactions. "The task", Norman wrote, must be "at just the proper level of difficulty: difficult enough to provide a challenge and require continued attention, but not so difficult that it invokes frustration and anxiety". A few years later (1999), Allan Cooper [22] introduced and discussed at length the concept of *cognitive friction*, defined as "the resistance encountered by a human intellect when it engages with a complex system of rules" and constraints imposed, for instance, by the technology they employ. Although Cooper primarily associated this concept with artifacts that are ill-designed and need to be improved to make them more usable, and thus as frictionless as possible, some measure of friction can be considered advantageous in light of Norman's insight: certain tasks should not be made too immediate because there is a risk of fostering attitudes of over-dependence, excessive complacency [58] and insufficient vigilance, as well as the risk of drifting users towards some form of deskilling [7].

In light of these insights, our paper presents an empirical comparative study that evaluates decision effectiveness and user experience in a system that *intentionally* introduces decision-making friction, as a way to foster more thoughtful and responsible human decision-making. We will delve deeper into this study and its background in the following section.

2 Motivations and background

The title of this work hints at a well-known scene from Star Wars (Episode V: The Empire Strikes Back) where Han Solo replies "Never tell me the odds!" to the anthropomorphous robot C-3PO telling him the (very low) odds of successfully navigating an asteroid field. Although this line can be considered a typical response to those who tell one what their chances of doing something are, we take it as a cue for an approach to AI development that requires these systems not to give odds, i.e., probabilities (or confidence scores), nor ready answers, clear-cut classifications, or predictions; but rather aimed at designing systems meeting the main requirements to help users think better [9].

Thus, the main motivation for this research grounds on the following conjecture: Decision support systems that provide full-fledged answers, such as the classification advice or quantitative or probabilistic estimates that the latest generation of AI systems can yield, might induce some form of over-dependence in users and, in the long run, a significant loss of skills (i.e., deskilling [64]) to their judgmental capacities.

While we are aware that this conjecture is still in need of strong empirical confirmation, we note that it has circulated in many environments where the

computer support of knowledge work, judgment, and case interpretation is most promising and effective (e.g., [15]), and a risk-based approach, such as that advocated by the EU regulation and other similarly prudent approaches, require that such a conjecture should be taken as true until proven otherwise.

Moreover, in contexts such as medicine, traditional systems can also encourage opportunistic behavior such as defensive medicine [34], i.e. deferring to the machine's answer to avoid accusations of negligence and malpractice. Paradoxically, this latter effect and the more general one of technology over-reliance [8] is all the more likely to happen the more accurate and reliable the systems are.

A possible preventive solution to this class of problems sometimes referred to as the unintended consequences of AI [7], less radical than abandoning decision support and foregoing its undoubted benefits, is to design supports that do not completely relieve the user of interpretative work, but rather promote it, by adding some *cognitive friction* to the decision-making process.

In this paper, we focus on a possible instance of *frictional AI*, which we will further characterize in Section 5: a solution that is based on the concept of a *pro-hoc explanation*. The name of this technique comes from its main feature to distinguish it from the more common *post-hoc explanations*: instead of providing the user with an explanation of the machine's answer after this has been given for a given case, like in case of post-hoc explanations, the system instead receives the user's tentative judgment (what in [11] is called human-first protocol and in [5, 33] *update cognitive forcing function*) and returns one (or more) possible explanations associated with that judgment (or counterfactual explanations for alternative outcomes). Therefore, instead of giving an explanation after a machine advice (post-hoc explanation), this solution entails the provision of a pro-hoc explanation (pro-hoc, from Latin, 'instead of that'), *which substitutes the machine advice*.

In this relatively unexplored domain, our focus is on *explanations by examples* [48], wherein the system presents cases akin to the current one. The concept of employing similar cases is not novel in AI research [68], and has been applied, for instance, in machine learning for similar image retrieval in medical contexts [21]: A prominent example is the SMILY system [37, 16]. However, our study diverges in both its objectives and scope from these precedents. Previous studies primarily concentrated on augmenting the image retrieval process itself [37], or on enhancing pathologist engagement to refine search outcomes based on visual similarity [16]. In contrast, our research ventures into decision support via pro-hoc explanations. We are not seeking to develop new algorithms for similar image retrieval or to explore how presenting comparable cases might bolster conventional predictive systems. Instead, our focus is on investigating the impact of presenting clinicians with analogous cases as the sole form of explanation (termed pro-hoc explanations above) for each potential outcome in a binary decision-making process.

Our unique contribution thus lies in examining the impact of these example-based pro-hoc explanations on the decision-making process in clinical settings, particularly in terms of cognitive effects and decision accuracy.

Specifically, in this paper, we will report the results of an exploratory user study in which (see Figure 1) we provided the users with similar cases depending on their prior judgment, both pro-hoc explanations that support the user's hypothesis, as well as counterfactual explanations responding to the objection "what if you were wrong? this would be the most similar case with the opposite label". In the latter case, we aim to see whether such "cognitively non-invasive" and "non-substitutive" support is effective (i.e. allows users to improve their baseline performance) and also perceived as such, i.e. useful, or not.

The case study above is designed in the context of radiological interpretation and diagnosis, for the task of identifying vertebral fractures from x-rays. In what follows, we will thoroughly describe the methods adopted to conduct the experiment (Section 3) and report the results therein collected (Section 4). Section 5 elaborates on the concept of frictional AI, and Section 6 discusses the study results and their implications. Finally, Section 7 concludes the work.

3 Methods

In what follows, we describe the methods applied to conduct the study to demonstrate whether giving physicians similar cases retrieved by the training set was decision-effective, i.e. it increased the users' diagnostic accuracy and was perceived as useful by them, even if this kind of support substituted traditional diagnostic support and abstained from classifying the new case.

In this experiment, we involved 16 physicians with varying degrees of experience in reading spine x-rays in their daily work, that is board-certified orthopaedic spine subspecialists (N=10) and orthopaedic residents (N=6). Their task was to annotate 18 x-rays cases, which had been selected in a previous study [9] for their representativeness of varied and complex cases, in terms of positive images (presenting some vertebral fracture) and negative images (with no vertebral fracture). The human-AI interaction protocol of this study, described in Figure 1, was kept as simple as possible: each orthopedist was presented one case at a time, through an online questionnaire, which had been implemented on the LimeSurvey platform [47]. For each case, each medical doctor was asked to provide a diagnostic opinion in terms of the presence (positive) or absence (negative) of lesions and fractures in an x-ray of 800x800 pixels (HD1 in Figure 1) and to indicate the perceived degree of difficulty (or complexity) of the case and his or her confidence in the proposed diagnosis, on a 6-value ordinal scale. Based on this first opinion (positive/negative), recorded by the system as HD1, the AI system retrieved, from the repository of available cases, the two most similar cases that presented (or did not present, respectively) fractures, as well as the most similar cases that did not have (or had, respectively) some fracture (see the middle step in Figure 1). The similar cases were retrieved according to their Cosine similarity, which was the similarity metric found to be more correlated with human ratings in a previous user study [13]. Conceptually, the experiment was the implementation of a *human-first* [11] (or second-opinion), partially *critiquing* [35] human-AI collaboration protocol. After considering these three similar cases, each physician had to indicate his or her final diagnosis, recorded by the system as FHD, also

indicating confidence in his or her final choice (see the last step in Figure 1). The physician's baseline accuracy, or pre-support accuracy, is then the observed success rate at the HD1 level; the post-support accuracy is FHD; the AI support regards the retrieval and visualization of similar cases, associated with their ground truth diagnosis, without proposing any categorical advice or probability scores, that is by abstaining from interpreting the case at hand: that is, we proposed the doctors to consider alternative *pro-hoc explanations* that support either their initial judgment or its opposite.

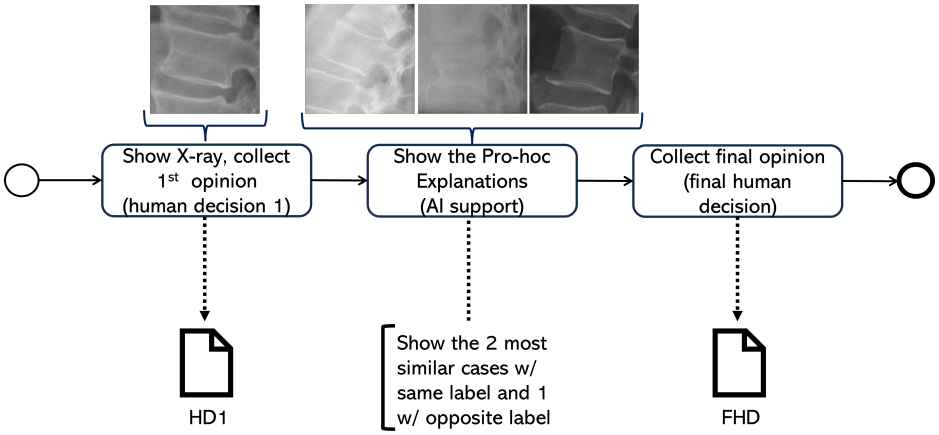

Fig. 1: BPMN diagram depicting the experimental design of the study. The x-rays are real cases, the examples regard two positive cases and one negative case.

We collected their 288 diagnoses and analyzed them by adopting a confidence level of 95% and applying non-parametric hypothesis testing. We also computed the so-called Number of Decisions Needed (NDN) to get a significant effect on the decision-making. This indicator was computed by Eq. (1), which is based on the Number Needed to Treat (NNT) used in epidemiology:

$$NDN = \frac{1}{(2 \times \text{pnorm}\left(\frac{d}{\sqrt{2}}\right) - 1)}, \tag{1}$$

where pnorm is the integral from $-\infty$ to q of the probability density function of the normal distribution and q is a Z-score, such as the effect size at hand. Intuitively, the NDN represents the average number of decisions users must make with the provided support (i.e., pro-hoc explanations) for one decision to be correct, as opposed to the likelihood of making incorrect decisions without such aid: in other words, the NDN is the number of decisions users must make with the given support to prevent an incorrect decision they would have otherwise made without the aid.

4 Results and main interpretations

In what follows, we report the results of the user study described above and outline some conjectures on the main factors that these results help to highlight.

What was the impact of showing similar cases instead of regular classifications (the pro-hoc approach) upon decision performance? As shown in Figures 2 and 3 it was small but positive: the pre-support accuracy of the participants was 78.8%, while their post-support accuracy was 80.9% (two proportion test p-value= .53, Z= -0.62, effect size .05). Although this results is not statistically significant, this small increase is better interpreted in the light of the *Number of Decisions Needed*. In fact, the observed NDN is 50, suggesting that using this system for approximately 50 decisions would suffice to avoid a mistake that would have been made without its adoption. Moreover, small effect sizes are typical in studies that evaluate the impact of XAI on diagnostic accuracy, when this impact is decoupled from the AI's effect (which is usually substantial) [8, 11]. For instance, in a similar setting considering 1548 diagnoses and 12 physicians, the effect size of providing explanations in the form of visual pixel-attribution maps associated with an 80% accurate support was 0.08 [10].

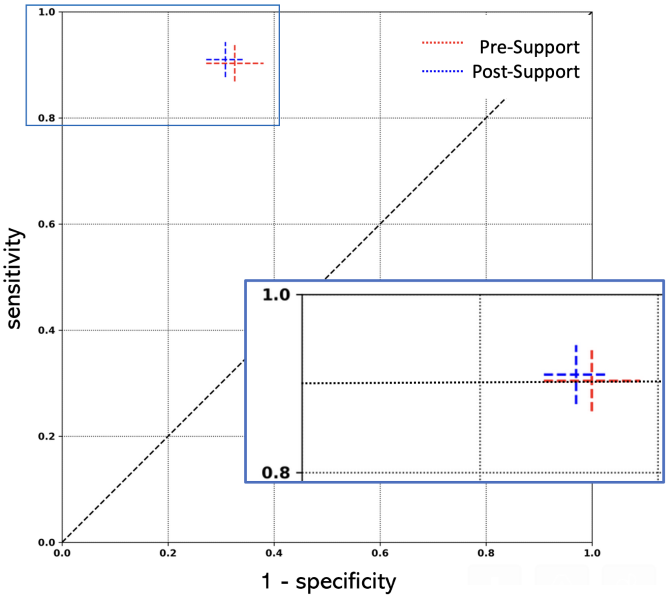

Fig. 2: Pre- and post-support average performances in the ROC space. Pre-support performance means sensitivity and specificity observed before showing similar cases, that is the pro-hoc explanations; post-support means performance exhibited by decision-makers after being exposed to the pro-hoc explanations.

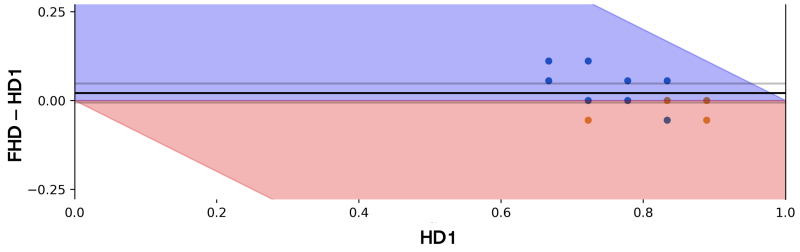

Fig. 3: Detail of a Benefit diagram illustrating the difference between pre and post-support average performances (generated with the online tool `https://mudilab.github.io/dss-quality-assessment/`). The blue region signifies performance enhancements attributed to the AI system, whereas the red region denotes performance decline resulting from its application. Blue dots represent specialist doctors, while orange dots correspond to resident doctors.

The small effect of showing similar cases on accuracy can be traced back to a very small number of decision changes (16 over a total of 288 decisions, see Table 1 which reports the reliance pattern [11] frequencies observed in the study). Notably, however, most decision changes were for the good: the number of decision changes from an initially wrong to a correct diagnosis (11) was more than double the number of decision changes in the reverse direction (5): this is why the observed effect of the pro-hoc explanations was found to be positive, as it accounted for a 10% reduction of diagnostic errors (see caption of Table 1).

Table 1: Reliance patterns' table for the user study. The Decision Support System (DSS) is the provision of Pro-Hoc explanations by similar cases. Zeros stand for wrong answers, while 1s for right ones. Mistakes passed from 61 (pre-DSS) to 55 (post-DSS): thus 6 mistakes were prevented by the tool, i.e., a 10% reduction. This table was generated by the online tool `https://mudilab.github.io/dss-quality-assessment/`.

Pre-DSS (HD1)	Post-DSS (FHD)	Count
0	0	50
0	1	11
1	0	5
1	1	222

The number of positive cases correctly identified, as well as the sensitivity and specificity, differed between the pre-support and post-support settings, but not significantly so (respectively, 177 vs 173, p-value=.73, Z=0.34, es=.03; .903 vs .910, p-value=.84, z=-0.20; es=.02, NDN=89; .674 vs .708, p-value=.52, Z=-0.64; es=.08, NDN=22). As a matter of fact, this latter not-so-slight increase in specificity could suggestively back up the finding, emerging also in other studies

(e.g.[29]), that explanations (as similar cases are) can make decision-makers less risk-averted (under the interpretation that the call for a 'negative test', which is more conservative for the patient, is more risky for the physician, in case they are wrong, in light of potential malpractice claims).

In regard to accuracy, we observed some interesting differences between residents (N=6) and specialists (N=10) (see Figure 3). When unaided, the residents performed better than the specialists, with an average accuracy of .83 (SD: .06) vs .76 (SD: .08) and with a large effect size (1.03, NDN = 2), albeit not significantly so (p-value= .5, T=2.1). This could be due to a greater commitment to the task of the residents (as compared with specialists), who took the opportunity to test their skills and (informally) compete with each other: indeed, on average, residents took 4 minutes more (15%) to complete the task than specialists.

However, this performance gap almost completely disappeared after the participants were shown similar cases. Indeed, although showing similar cases has improved the physicians' average accuracy in a not significant way (the p-value equalling .41, the test statistic T equalling -0.84, with a mean accuracy across the physicians that changed from .79 [.74,.82], .81 [.77, .84]), the observed effect size for the specialists was small-to-moderate (.3, NDN = 6). In particular, while 60% of them improved their accuracy, no resident improved their (see the Benefit diagram in Figure 3). This could be due to fixation by the residents. Indeed, when aided by the system, specialists changed their minds twice as frequently as residents (4% vs 7%): in two-thirds of these changes, the decisions regarded the diagnosis of cases that were deemed to be complex (that is the cases whose perceived complexity was evaluated higher than 2). More notably, two-thirds of these decision changes were for the better, whereas one-third induced some form of automation bias: nevertheless, automation bias was two and a half times larger in residents than in specialists (0.028 vs. 011). Most notably the rate with which specialists changed their minds for the better was 6 times greater than the rate for residents (5.6% vs 9%).

Although the positive effect on accuracy was found not to be significant (for the small sample of decisions considered), as we already commented on above, the observed effect sizes, especially for the specialists, were not negligible, nor the NDN: the system helped to avoid a potential mistake every 6 aided diagnoses. This makes us conclude that showing similar cases had a positive effect on the radiological task considered, and could therefore be considered useful: this finding is also confirmed by noting the Technology Impact, TI (see Figure 4). This is a measure of the usefulness of AI support, which was introduced in [8] and is defined as the ratio of the probability of making a correct decision when supported to the probability of making a correct decision when unsupported. For this study, the TI was slightly positive and significantly so: as shown in Figure 4, the 95% confidence interval for the odds ratio does not contain the line of 'no impact' (TI=1).

To confirm this statistical finding from a more qualitative perspective, we also asked directly to the participants in the case study if they found the support useful. Not surprisingly then, the perceived usefulness (evaluated for every single decision on a 4-value ordinal scale) was high (average: 2.7, 95% confidence interval

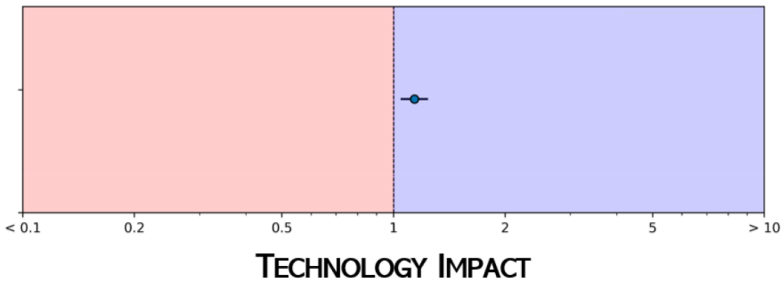

Fig. 4: Technology Impact odds ratio, for the decision support study. Horizontal lines denote 95% C.I. computed according to the standard formula for odds ratios. The red region denotes an overall negative effect of the AI support, while the blue region denotes an overall positive effect (moreover, since the confidence interval does not cross the boundary, the effect is significant). Diagram generated with the online tool `https://mudilab.github.io/dss-quality-assessment/`.

[2.62, 2.87]) and the vast majority of respondents chose a value in the upper half of the scale (.62 vs .38, significant majority with p-value from the binomial test .0002). The difference in perceived usefulness between residents and specialists is significant (p-value= .010, Mann Whitney Z= 2.57, standardized effect size= .16, common language effect size= .59). As it can be seen in Figure 5, residents considered the aid more useful than the specialists, although its impact on their accuracy, and hence the augmentation effect, was much smaller, as also confirmed by the Benefit Diagram depicted in Figure 3.

Considering similar cases made respondents slightly more confident (but not significantly so) with respect to not having any support. The p-value equals .176, the test statistic Z equals -1.353. The observed standardized effect size is small (0.056, NDN = 32), but the observed common language effect size was moderate (0.47). For specialists, the difference was stronger, although still not significant (P= .07724) with a standardized effect size almost twofold bigger (.093, NDN = 19). More generally, confidence changed in slightly more than 55% of the cases for which physicians changed their decision, and within these cases, confidence improved in two-thirds of the decision changes. Indeed, after seeing similar cases, the number of times the physicians' confidence increased is significantly higher than the number of decreases (.66 vs .34, p-value=.0014; test statistic X = 64) and effect size is .34 (NDN = 5). Thus, although in half of the cases reported confidence did not change, in almost one-quarter of the cases presenting similar cases improved the confidence of decision-makers in their decisions. On the one hand, this does not surprise, as similar cases are additional information that, ideally, can make physicians more certain of their diagnosis; on the other hand, this result suggests that similar cases are considered useful in

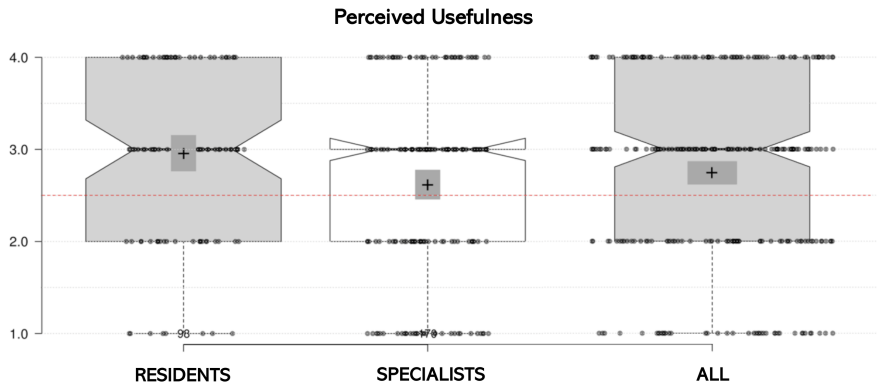

Fig. 5: Boxplots of the perceived usefulness of showing similar cases for each diagnostic decision (N=288), reported by residents (on the left) and specialists (on the right) on a 4-point ordinal scale. Box notches represent 95% confidence intervals of the median; crosses represent means, in their 95% CIs as well. The horizontal red line represents the mid-scale: estimated averages that do not cross this line indicate statistically significant trends in the collected responses.

corroborating confidence about the correctness of one's diagnosis, and this effect could not be given for granted.

Finally, we also observed some interesting correlations between the observed accuracy of the physicians and their perceptions (see Figure 6) as well as between these perceptions and the perceived complexity of cases (see Figure 7). More in particular, we notice that the psychometric variables reported by the participants, that is their confidence and the perceived complexity of the cases, were a reliable proxy of their actual accuracy. As expected, the more confident the physicians were about their diagnosis (HD1) in regard to a case, the higher the actual accuracy (see Figure 6, on the left), that is the match between FHD and the ground truth. Likewise, the higher the perceived complexity, the lower the actual accuracy (see Figure 6, on the right). In both cases, correlation scores were high and significant (resp, confidence: +.48; complexity: -.39). We observed also two other significant and strong correlations: between the perceived complexity of cases and confidence in the final decision (see Figure 7 on the left), and between perceived complexity and perceived usefulness of pro-hoc explanations (see Figure 7 on the right). As understandable, the higher the perceived complexity, the lower the confidence (correlation: -.78) and, most notably, the higher the perceived usefulness of the AI support (+.32). These results confirm the ecological validity of our study and the importance of involving real subject matter experts in Human-Centered Artificial Intelligence and Medical AI studies.

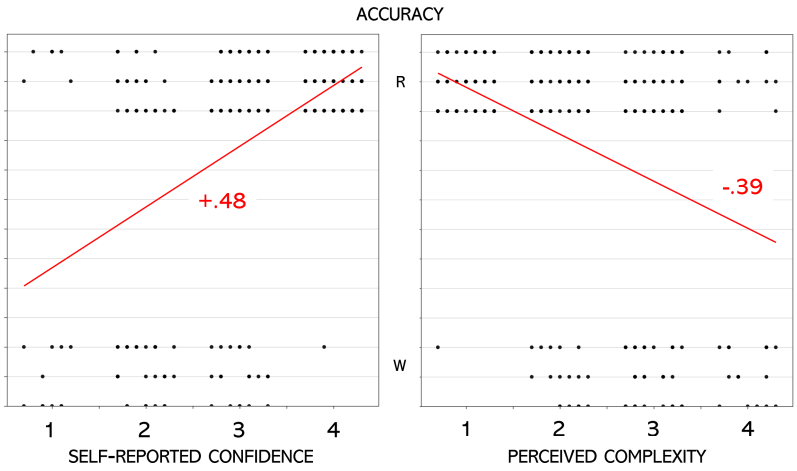

Fig. 6: Scatterplots of the relationship between self-reported confidence in the decision, on the left, and reported perceived case complexity, on the right, with actual accuracy (vertical axis). Points correspond to single decisions, jittered to avoid clutter and overlapping. Pearson correlation coefficients are indicated in red, close to the regression trend.

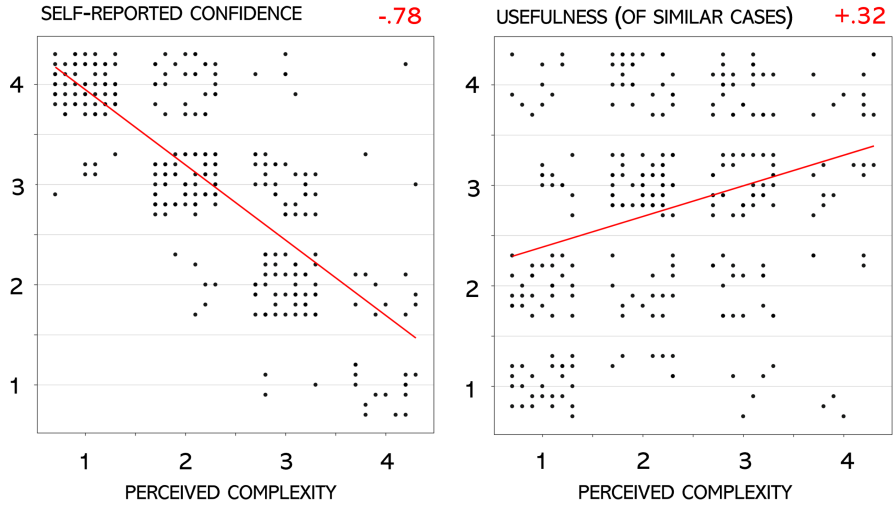

Fig. 7: Scatterplots showing the correlations between case (perceived) complexity and, respectively, the physicians' confidence on their initial diagnosis (left), the physicians' accuracy (middle), and the perceived usefulness of the 'similar case' support. Spearman correlation coefficients (reported in red) are all statistically significant.

5 Frictional AI

Before delving into the implications of our findings, it is crucial to elaborate on the concept of frictional AI, briefly introduced in Section 1: this discussion will provide a framework for better appreciating the scope and impact of our results.

As mentioned earlier, the term 'cognitive friction,' as initially presented by Cooper in 1999 [22], has predominantly been viewed as an inadvertent consequence of poor design in interactive systems, rather than as a potential tool to encourage more mindful user engagement in the spirit of Norman [56].

The promotion of a more mindful approach to technology echoes Kahneman's seminal work "Thinking, Fast and Slow" [41], which distinguishes two main modalities of human thinking: System 1 and System 2. System 1 represents our fast, intuitive, and "automatic" decision-making, while System 2 concerns slower, analytic, rational thinking. Despite being a simplistic interpretation of human thinking [44], the two-system framework can help us think of over-reliance not as an intrinsic, unavoidable phenomenon due to cognitive biases, but rather as a specifically designed mode of human-AI interaction in which a Decision Support System (DSS) leverages System 1 more than System 2, usually for the sake of efficiency. Conversely, to create the best conditions facilitating commitment, oversight, and responsible caution, we investigate AI systems that elicit System 2 thinking through careful interaction designs that disrupt or discourage automatic user behavior. [19, 24, 51, 54].

One such example regards systems that embed micro boundaries [24]: these are defined as small obstacles or moments of reflection that "create just enough friction to switch someone from having their behavior driven by System 1 to System 2" [24] by slowing down the decision-making process. In this sense, micro boundaries are in stark contrast with so-called dark patterns, namely design choices aimed at "sludging" users into undesired behaviors by exploiting their inattentiveness and eliciting quick, instinctual responses [32, 24]. Dark patterns produce an intentionally seamless and smooth experience of use, to conceal complexity and secondary motives. A completely different approach, called seamful design, intentionally reveals system shortcomings and the "mismatches and cracks between assumptions made in designing and developing the AI system and the reality of the deployment context" [28], to promote a balanced level of user reliance [38], and make users more aware of existing uncertainty and inconsistency [40].

Thus, in the wake of the seminal work by Norman [56], some researchers began to consider the appropriateness of including elements that *intentionally* cause friction, under the names of critical design [27], reflective design [66], adversarial design [25] and the concept of 'intentional - beneficial - friction' [74]. The cognitive friction that these approaches envision can be rendered in multiple ways: for instance, by disabling functionalities that might otherwise be expected or desired; by making them more difficult to run; or by purposedly introducing slow-downs, pauses, and inefficiencies.

Frischmann and Selinger [30] convincingly argued that "some friction, some inefficiency, even some transaction costs may be necessary to sustain an underde-

termined environment conducive to human flourishing." (p. 141). From the same authors, this idea was then concretely translated in terms of *programmed inefficiencies*, which are "deliberately engineered" "sources of friction" (p. 286) [30], that is, design features that implement what Ohm and Frankle [57] called one year earlier "desirable inefficiency" (i.e., a design pattern that connects apparently inefficient code and human values), and a little earlier the authors of [70] called "inspired inefficiency" (i.e., the result of balancing algorithms and intuition), and the authors of [31], probably predating all the others, called "meaningful inefficiencies" (p. 254) in the context of civic life. We also conducted some empirical studies on the concept of programmed inefficiency [7, 12] in medical decision-making and second-opinion settings. Known precursors of similar approaches are slow technology [36] and reflective design [66], which focus on how technology can encourage and aid a thoughtful and considerate demeanor in users throughout the interaction. Conceptually preliminary proposals in that direction are "uncomfortable interactions" [3] and "critical design" promoting reflection and critique through making technology "unfriendly" to users [26] or subverting assumptions and expectations, like the strong one that decision support systems should only give recommendations and pieces of advice.

Pierce introduced the idea of integrating "digital limitations" when designing "counterfunctional things" (that is "a thing that figuratively counters some of its own functionality"), partly under the influence of the tenets of nudging theory and the research on choice overload. More recently, Pierce presented a framework for "frictional design", which grounds on his pioneering research on undesign [60] and alternative designs, which include five tendencies: "diverging, opposing, accelerating, counterfactualizing, and analogizing" [61].

Acknowledging Pierce's contribution [60], we introduce Frictional AI as the umbrella term for a set of various approaches whose aim is to design AI systems that promote reflection and critique rather than complacency or mindless reliance. This approach carefully inserts design frictions [24] or programmed inefficiencies [7] instead of recklessly removing them to make interaction faster and more efficient. This is done following the idea that it is designers, rather than users, who are to be held responsible for creating the best conditions facilitating commitment, oversight, and responsible judgment.

In the domain of DSS and AI design, reflection machines [23] are systems that prompt users to critically reflect on their own decision-making strategies; evaluative AI [53], on the other hand, denote systems that do not offer direct recommendations but rather provide evidence for and against a specific decision. Both these approaches can be considered as instances of what we call frictional AI. Also in the case of pro-hoc explanations, the friction is intended to mitigate the risk that users might over-rely on the support and develop heuristics or opportunistic behaviors in which they exert low vigilance or overtrust in the system advice, even when this is wrong: automation bias and complacency are the terms usually associated with these behaviors when they regard individual decisions and choices [58, 73], but also the risk of deskilling has been reported in the long run [64].

Thus, we call frictional AI the composite field of design research aimed at applying the above tenets and insights to the development of AI systems and data-driven decision support systems, in order to create some cognitive friction in the human process of situation assessment or decision-making. In what follows we provide a typology of frictional AI applications, expanding on the work by [55], without any ambition of exhaustiveness and completeness.

1. **Cautious protocols**, where the system presents multiple options or none. In the former case, the system presents a set of candidate answers, which are associated with either an individual confidence score each, or with a defined level of probability of encompassing the right answer, like in conformal prediction. The latter case is what is also called abstention, that is the deliberate rejection to provide support, or a a degenerate case of conformal prediction in which all possible options are mentioned. In all those cases, the system recognizes the case as being too complex or too different to the cases seen in the training set [17], and applied this kind of protocol not to mislead the decision-maker.
2. **Judicial or antagonist protocols**, where the system hosts arguments and explanations backing up multiple and opposite decisions or interpretations. A particular case envisioned in [53] is that of perorative explanations produced by opposing conversational agents, which try to convince the human decision-maker that their interpretation and classification is the right one, while the other is wrong. Another instance is that of agonistic machine learning models (two or more) that provide opposite answers and related explanations. These models could belong to different model families or apply different hyperparameters, or be trained on different ground truths and representations [39] or be optimized for different targets such as utility, specificity, sensitivity or discriminative performance [49]. The introduction of such "conflicting rules/knowledge" is what Kliegr [45] identified as a debiasing technique against *overconfidence* and *underconfidence* and has been previously studied by Wang et al. [72], Bhatt et al. [4], Bussone [6], and Wolfe [75].
3. **Decentralized AI or adjunct protocols**. These protocols, first introduced in [14] with the term *adjunct AI*, employ *process friction* to make decision-making less dependent on AI or to make AI-supported processes less effective than unaided ones. This can be accomplished, for instance, by embedding cognitive forcing functions [5] such as timeout periods, longer waiting times, and purposefully slowed-down algorithms, which have been found to improve user evaluation of algorithmic accuracy [59]. Another case of process friction entails assigning the AI to the role of a second-opinion giver [69], after that the human decision-maker has recorded their first opinion. This solution has been applied in [11] and denoted as *human-first* protocol. Both solutions are aimed at mitigating biases, such as algorithmic deference, selective adherence, priming effects, framing, and anchoring bias [1, 62, 63], as the user is required to come up with their own interpretation on the case at hand before being influenced by the AI output. Thus, adjunct AI protocols give value to human intuition and aim to complement it, rather than substitute it. In our user study, we applied this kind of protocol, combined with the following one;

4. **Comparative or analogical protocols**, where the system provides users with access to the most similar cases (to the case at hand) that are associated with their ground truth; or the most similar cases (to the case at hand) that are associated with each of all the available classes. Users are then invited to reflect on the elements that differentiate or liken the present case and past ones, orienting their final decision according to the labels associated with the previous cases [2]. In these cases, the system can be considered as a case-mining tool or transactive memory, rather than an *oracular* [52] support, which fosters analogical thinking [42]. This is exactly the case of *pro-hoc* explanations, which replace, rather than complement, the AI decision support.

6 Design implications and further considerations

This study aims to see whether showing similar cases is an effective alternative for giving explicit categorical advice, that is testing the effectiveness (and usability) of *pro-hoc explanations*. The main findings are summarized in Table 2: they suggest that *pro-hoc explanations by examples* are considered useful (see Figure 5) and they might also increase accuracy, although minimally so, since the observed effect was very low (see Figures 2 and 3). However, a relatively low number of decisions are necessary before avoiding some mistakes that would be committed without the support, as shown by the NDN indicator. This finding, as well as the fact that showing multiple similar cases (with known and verified diagnoses) is tantamount to not giving physicians any ready-made answer, suggest that pro-hoc explanations can be considered one of the Frictional AI solutions associated with the smallest risk of deskilling and over-reliance: indeed, this kind of solution still require physicians to exert their interpretative skills and judge the images by themselves.

All things considered, from the findings above (see also Table 2), we can draw some guidelines and recommendations. We outline them in what follows, with no particular ambition of generalization, but with the aim to inform the design of frictional AI systems aimed at improving diagnostic accuracy in radiological settings, as well as to stimulate further research in this domain:

- In image annotation and ground truthing, information about case/decision complexity and diagnostic confidence should be collected: indeed, as we have shown in our experiments (see Figure 6), these data correlate with actual accuracy. Moreover, this information could be useful to modulate the level of friction of the system, as seen in [18].
- Showing cases similar to the one under examination, including both negative and positive instances of a specific condition (e.g., vertebral fractures), may assist readers in their diagnostics. This potential improvement appears to be more perceptible among less experienced readers than expert ones.
- The presentation of similar cases may enhance the confidence of image readers in their final decisions. This effect could be more pronounced among expert readers, though it is not exclusive to them. It's important to note that this

Table 2: Summative table of the main findings and the related implications for either the development of the XAI feature at hand or the design of future empirical studies evaluating its effectiveness. The Number of Decisions Needed (NDN) is computed according to Eq. 1. The Recommended Sample Size (RSS), in terms of the minimum number of decisions to observe to get statistically significant results, has been produced with the procedure presented in [67], and by adopting a Power of .8 and α of .05.

Finding	**Effect Size**	**Number of Decisions Needed (NDN)**	**Recommended Sample Size (RSS)**
Pro-Hoc Explanations improve user diagnostic accuracy	.05	50	6000
Pro-Hoc Explanations improve expert diagnostic accuracy	.30	6	320
Pro-Hoc Explanations make users more confident	.06	32	3500
Pro-Hoc Explanations make experts more confident	.09	19	2800

increased confidence may occur even if there is no change in their final decision (i.e., when $HD1 = FHD$). While not definitive, this trend towards greater confidence after using the DSS might be interpreted as a form of satisfaction or an indicator of a positive user experience. Consequently, this suggests that pro-hoc explanations could potentially enhance the usability of the DSS, although this effect may vary among users.

Since the last point concerns one of the most important elements for those involved in the HCI field, we investigated this point in greater detail, by involving the two authors who are expert clinicians and who were originally involved in the design of the *pro-hoc* functionality. They both found the adoption of this functionality intriguing to use and of high potential in terms of usability: indeed, they noticed that learning from analogous cases reported and described in scientific articles, as well as in textbooks or congress presentations, is a common and essential aspect of medical education [20, 65]. As a consequence, clinicians are very familiar with this kind of unobtrusive aid: this could be a factor in regard to the rise in confidence that we observed in this study, even in those cases where the physicians did not change their initial diagnostic interpretation.

To better illustrate how a clinician could leverage the above functionality and trigger effective analogical reasoning, or at least a line of reasoning that increases

their confidence and hence satisfaction, we discussed with the above clinicians some cases used in our study in greater detail.

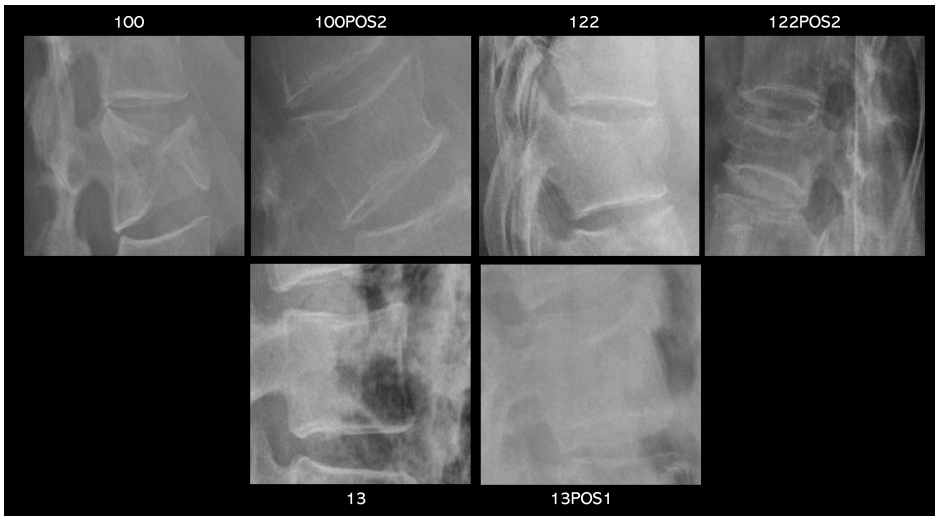

Fig. 8: Cases from the experimental set where the AI aid was proved to be useful in either confidence or accuracy improvements, in virtue of the similarity between the case at hand and those retrieved.

For instance, case 100 (see Figure 8) was associated with highly correlated similarity scores with those retrieved by the AI. One of these retrieved cases (see case 100POS2 in Figure 8) exhibited the same type of fracture as the former case index, which nevertheless is extremely rare (A2, in the AOSpine classification), as it is observed in only 3% of fractures [71]. For its rarity and appearance, case 100 (and similar cases) are relatively easy to diagnose (for expert readers) and hence the perceived utility of the AI aid was relatively low, although it was effective in retrieving a conceptually similar case.

In other cases, where diagnosing a fracture was less straightforward, similar cases were perceived as more useful. For instance, in regard to cases 122 and 13 (both depicted in Figure 8), which are actually hard to diagnose, the AI retrieved two similar cases (respectively, case 122POS2 and 13POS1 in Figure 8) that presented the same specific fracture pattern and anatomical lesion (i.e., anterosuperior corner fracture) and had been associated with a verified diagnosis of fracture presence, thus suggesting that also the former case (122POS2) was positive. Consequently, not only the AI aid did improve the physician's confidence in the final diagnosis, but it also provided a differential benefit in regard to accuracy, and the more so especially for those cases whose difficulty level was higher (i.e., MIO grading 3 and 4).

However, not all decisions were so straightforward and some cases posed interpretative challenges. For instance, case 65 (see Figure 9) exhibited a fracture

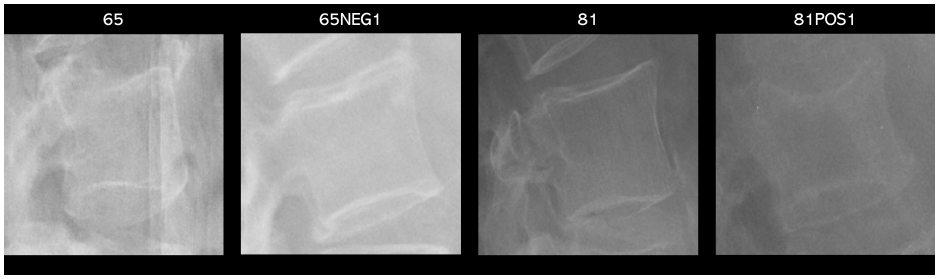

Fig. 9: Cases from the experimental set where the AI aid was controversial or potentially misleading, in virtue of the similarity between the case at hand and those retrieved.

with a lesion pattern that is hard to detect and recognize on plain X-rays, usually resulting in lower detection rates and higher error rates. In a real-world and naturalistic setting, any fracture diagnosis for a case like that would have to be confirmed using second-level imaging, such as a (much more expensive and invasive) CT or a much more stressful and expensive exam such as MRI. For the above case, the more similar cases retrieved by the AI had been labeled as negatives (e.g., see case 65NEG1 in Figure 9), thus potentially confusing and potentially misleading the clinician. Similarly, case 81 presented a comparable scenario with a different initial diagnosis. In this case, the case was actually negative for fracture (according to the ground truth), but the most similar case retrieved by the AI had been labeled as positive for fracture (see case 81POS1 in Figure 9). This could also mislead the physician, who might have been led to believe that they were missing a fracture diagnosis and that a second (unnecessary) imaging was to be prescribed. Nevertheless, it is worth noting that from a clinical perspective, the impact of a missed fracture is much more serious than that of an incorrect diagnosis of a fracture. In the former scenario, the patient could go untreated, resulting in potentially serious consequences [43], whereas in the latter scenario, second-level imaging, or further physical inspection, would lead to the correct diagnosis. Therefore, in orthopedic settings, an AI aid should be designed to optimize sensitivity (over specificity), that is to make false negatives more rare to result in a more useful support [18]. The above comments choose to adopt a pro-hoc support a design choice that can have different effects based on contextual elements, such as the complexity, difficulty or rarity of the case: this an additional reason to collect the physicians' perception at use time, to classify cases also along these subjective dimensions.

7 Conclusions

While friction is typically seen as a drawback in HCI, increasing research indicates its potential to foster more deliberate, mindful, and critical interactions with digital systems. In this paper, we contribute to the research that explores the deliberate incorporation of friction in decision support systems. This approach

aims to encourage more conscientious decisions and counterbalance the risks of excessive reliance on technology and its dominance, in the short term [8], and to address concerns about skill deterioration and learning degradation, in the long term [7].

The design and evaluation of human-AI interaction protocols are inherently human-centered because they are built grounding on users' perceptions and involve direct user participation. Therefore, our research aligns with both the Human-Centered Artificial Intelligence and the *One Health* approaches for two main reasons: firstly, our user study aims to improve the efficacy of clinical DSSs and, by extension, medical decision-making, although the insights gained may also be applicable to other decision-making contexts. Secondly, we embrace the core principle of the *One Health* initiative, which posits that technologies intended to ensure the health of individuals and their environment should be designed holistically, taking into account long-term effects and distant externalities associated with any improvement in human decision performance. In this context, we consider the innate variability of human beings and their propensity for minimal effort, shortcuts, and 'cognitive economy' strategies [50], as motivations to introduce controlled friction in decision-making processes and assess their potential in reducing over-reliance on technology, technology dominance [8] and the tendency to perceive AI systems as agents that are equally or more capable than humans, rather than as mere tools that enhance our cognitive abilities [12].

To this aim, we conducted an empirical user study in a controlled environment for the diagnostic task of detecting vertebral fractures in spine x-rays. The interaction protocol, as illustrated in Figure 1, prioritized human judgment: clinicians first made an initial assessment of the x-ray without AI assistance and then provided a final diagnosis with AI support. This support involved displaying three similar cases identified by the AI: two matching the physician's initial classification and one from the opposite class. While the application of this kind of support only marginally improved accuracy (approximately by an additional 2%), its utility was greatly appreciated by the medical practitioners involved, particularly those with less experience, despite more experienced clinicians showing greater improvement. Participants also reported increased confidence in their final decisions.

A key limitation of this study is the small scale, both in case numbers and participant count. While this restricts the generalizability of our findings, the identified effects are substantial enough to inform future research design: specifically, the identified effect sizes are of critical importance to support the power analysis and estimation of the sample size of future studies (see Table 2). This makes our study capable of informing the design of future, more-powered, studies involving potentially larger, and more diverse, samples of clinicians.

Our future work will also go in this above direction: more specifically, we plan to further explore how frictional AI protocols, like pro-hoc explanations, can support human users, particularly in preventing errors. As discussed in Section 6, pro-hoc explanations help users by allowing them to compare their preliminary assessments against actual diagnoses of closely similar cases, thereby prompting a reassessment of their initial reasoning and conclusions. The promising results

reported in this study suggest that further research should focus on evaluating other methods by which less immediately exploitable AI outputs can still augment human decision-making, according to the typology outlined in Section 5.

Acknowledgements

F. Cabitza acknowledges 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. Alon-Barkat, S., Busuioc, M.: Human–ai interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. Journal of Public Administration Research and Theory **33**(1), 153–169 (2023). https://doi.org/10.1093/jopart/muac007
2. Baselli, G., Codari, M., Sardanelli, F.: Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? European Radiology Experimental **4**, 1–7 (2020). https://doi.org/10.1186/s41747-020-00159-0
3. Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J., Rodden, T.: Uncomfortable interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 2005–2014. CHI '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2207676.2208347
4. Bhatt, U., Antorán, J., Zhang, Y., Liao, Q.V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al.: Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 401–413 (2021). https://doi.org/10.1145/3461702.3462571
5. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: 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**(CSCW1), 1–21 (2021). https://doi.org/10.1145/3449287
6. Bussone, A., Stumpf, S., O'Sullivan, D.: The role of explanations on trust and reliance in clinical decision support systems. In: 2015 international conference on healthcare informatics. pp. 160–169. IEEE (2015). https://doi.org/10.1109/ICHI.2015.26
7. Cabitza, F.: Cobra AI: Exploring some unintended consequences of our most powerful technology. Machines We Trust: Perspectives on Dependable AI **87** (2021). https://doi.org/10.7551/mitpress/12186.003.0011
8. Cabitza, F., Campagner, A., Angius, R., Natali, C., Reverberi, C.: 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. CHI '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581095
9. Cabitza, F., Campagner, A., Famiglini, L., Natali, C., Caccavella, V., Gallazzi, E.: Let me think! investigating the effect of explanations feeding doubts about the AI advice. In: International Cross-Domain Conference for Machine Learning and

Knowledge Extraction. pp. 155–169. Springer (2023). https://doi.org/10.1007/978-3-031-40837-3-10
10. Cabitza, F., Campagner, A., Natali, C., Parimbelli, E., Ronzio, L., Cameli, M.: Painting the black box white: experimental findings from applying xai to an ecg reading setting. Machine Learning and Knowledge Extraction **5**(1), 269–286 (2023). https://doi.org/10.3390/make5010017
11. Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G.E., Pastore, M.C., Sconfienza, L., Folgado, D., Barandas, M., Gamboa, H.: Rams, hounds and white boxes: Investigating human-ai collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine p. 102506 (2023). https://doi.org/10.1016/j.artmed.2023.102506
12. Cabitza, F., Campagner, A., Simone, C.: The need to move away from agential-ai: Empirical investigations, useful concepts and open issues. International Journal of Human-Computer Studies **155**, 102696 (2021). https://doi.org/10.1016/j.ijhcs.2021.102696
13. Cabitza, F., Famiglini, L., Campagner, A., Caccavella, V., Gallazzi, E., Sconfienza, L.M.: Similar to what? investigating the usefulness of retrieving similar cases with AI in radiological settings. In: submitted (2024)
14. Cabitza, F., Natali, C.: Open, multiple, adjunct. decision support at the time of relational ai. In: HHAI2022: Augmenting Human Intellect, pp. 243–245. IOS Press (2022). https://doi.org/10.3233/FAIA220204
15. Cabitza, F., Rasoini, R., Gensini, G.F.: Unintended consequences of machine learning in medicine. Jama **318**(6), 517–518 (2017). https://doi.org/10.1001/jama.2017.7797
16. Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., Terry, M.: Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–14. CHI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300234, `https://doi.org/10.1145/3290605.3300234`
17. Campagner, A., Cabitza, F., Ciucci, D.: Three-way decision for handling uncertainty in machine learning: A narrative review. In: Rough Sets: International Joint Conference, IJCRS 2020, Havana, Cuba, June 29–July 3, 2020, Proceedings. pp. 137–152. Springer (2020). https://doi.org/10.1007/978-3-030-52705-1-10
18. Campagner, A., Sternini, F., Cabitza, F.: Decisions are not all equal—introducing a utility metric based on case-wise raters' perceptions. Computer Methods and Programs in Biomedicine **221**, 106930 (2022). https://doi.org/10.1016/j.cmpb.2022.106930
19. Caraban, A., Karapanos, E., Gonçalves, D., Campos, P.: 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–15 (2019). https://doi.org/10.1145/3290605.3300733
20. Challa, K.T., Sayed, A., Acharya, Y.: Modern techniques of teaching and learning in medical education: a descriptive literature review. MedEdPublish **10**, 18 (2021). https://doi.org/10.15694/mep.2021.000018.1
21. Choudhury, N., Begum, S.A.: A survey on case-based reasoning in medicine. International Journal of Advanced Computer Science and Applications **7**(8), 136–144 (2016)
22. Cooper, A.: The Inmates are Running the Asylum. Vieweg+Teubner Verlag (1999)

23. Cornelissen, N., van Eerdt, R., Schraffenberger, H., Haselager, W.F.: Reflection machines: increasing meaningful human control over decision support systems. Ethics and Information Technology **24**(2), 19 (2022). https://doi.org/10.1007/s10676-022-09645-y
24. Cox, A.L., Gould, S.J., Cecchinato, M.E., Iacovides, I., Renfree, I.: Design frictions for mindful interactions: The case for microboundaries. In: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems. pp. 1389–1397 (2016). https://doi.org/10.1145/2851581.2892410
25. DiSalvo, C.: Adversarial design as inquiry and practice. MIT Press (2012)
26. Dunne, A.: Hertzian tales: Electronic products, aesthetic experience, and critical design. MIT press (2008)
27. Dunne, A., Raby, F.: Design noir: The secret life of electronic objects. Springer Science & Business Media (2001)
28. Ehsan, U., Liao, Q.V., Passi, S., Riedl, M.O., Daume III, H.: Seamful xai: Operationalizing seamful design in explainable ai. arXiv preprint arXiv:2211.06753 (2022)
29. Estepa-Mohedano, L., Espinosa, M.P.: Comparing risk elicitation in lotteries with visual or contextual aids. Journal of Behavioral and Experimental Economics p. 101974 (2022). https://doi.org/10.1016/j.socec.2022.101974
30. Frischmann, B., Selinger, E.: Re-engineering humanity. Cambridge University Press (2018)
31. Gordon, E., et al.: Civic engagement. Dialogues on mobile communication pp. 156–171 (2016)
32. Gray, C.M., Kou, Y., Battles, B., Hoggatt, J., Toombs, A.L.: The dark (patterns) side of ux design. In: Proceedings of the 2018 CHI conference on human factors in computing systems. pp. 1–14 (2018). https://doi.org/10.1145/3173574.3174108
33. Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–24 (2019). https://doi.org/10.1145/3359152
34. Grote, T., Berens, P.: How competitors become collaborators—bridging the gap (s) between machine learning algorithms and clinicians. Bioethics **36**(2), 134–142 (2022). https://doi.org/10.1111/bioe.12957
35. Guerlain, S.A., Smith, P.J., Obradovich, J.H., Rudmann, S., Strohm, P., Smith, J.W., Svirbely, J., Sachs, L.: Interactive critiquing as a form of decision support: An empirical evaluation. Human factors **41**(1), 72–89 (1999). https://doi.org/10.1518/001872099779577363
36. Hallnäs, L., Redström, J.: Slow technology–designing for reflection. Personal and ubiquitous computing **5**, 201–212 (2001)
37. Hegde, N., Hipp, J.D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C.J., Amin, M.B., Mermel, C.H., et al.: Similar image search for histopathology: Smily. NPJ digital medicine **2**(1), 56 (2019)
38. Hengesbach, N.: Undoing seamlessness: Exploring seams for critical visualization. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–7 (2022). https://doi.org/10.1145/3491101.3519703
39. Hildebrandt, M.: Algorithmic regulation and the rule of law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2128), 20170355 (2018)
40. Inman, S., Ribes, D.: "beautiful seams" strategic revelations and concealments. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2019). https://doi.org/10.1145/3290605.3300508
41. Kahneman, D.: Thinking, fast and slow. Farrar, Straus and Giroux (2011)

42. Keane, M.: Analogical mechanisms. Artificial Intelligence Review **2**(4), 229–251 (1988). https://doi.org/10.1007/BF00138817
43. Khatri, K., Farooque, K., Sharma, V., Gupta, B., Gamanagatti, S.: Neglected thoraco lumbar traumatic spine injuries. Asian Spine Journal **10**(4), 678 (2016). https://doi.org/10.4184/asj.2016.10.4.678
44. Klein, G.A., Orasanu, J., Calderwood, R., Zsambok, C.E., et al.: Decision making in action: Models and methods, vol. 3. Ablex Norwood, NJ (1993)
45. Kliegr, T., Bahník, Š., Fürnkranz, J.: A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artificial Intelligence **295**, 103458 (2021). https://doi.org/10.1016/j.artint.2021.103458
46. Krug, S.: Don't make me think!: a common sense approach to Web usability. Pearson Education India (2000)
47. LimeSurvey Project Team / Carsten Schmitz: LimeSurvey: An Open Source survey tool. LimeSurvey Project, Hamburg, Germany (2012), `http://www.limesurvey.org`
48. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018). https://doi.org/10.1145/3236386.3241340
49. Lu, D., Tao, C., Chen, J., Li, F., Guo, F., Carin, L.: Reconsidering generative objectives for counterfactual reasoning. Advances in Neural Information Processing Systems **33**, 21539–21553 (2020)
50. Lyell, D., Magrabi, F., Coiera, E.: The effect of cognitive load and task complexity on automation bias in electronic prescribing. Human Factors **60**(7), 1008–1021 (2018). https://doi.org/10.1177/0018720818781224
51. Mejtoft, T., Hale, S., Söderström, U.: Design friction. In: Proceedings of the 31st European Conference on Cognitive Ergonomics. pp. 41–44 (2019). https://doi.org/10.1145/3335082.3335106
52. Miller, R., Masarie Jr, F.: The demise of the "greek oracle" model for medical diagnostic systems. Methods of information in medicine **29**(01), 1–2 (1990). https://doi.org/10.1055/s-0038-1634767
53. Miller, T.: Explainable AI is dead, long live explainable AI! hypothesis-driven decision support. arXiv preprint arXiv:2302.12389 (2023)
54. Naiseh, M., Al-Mansoori, R.S., Al-Thani, D., Jiang, N., Ali, R.: Nudging through friction: An approach for calibrating trust in explainable AI. In: 2021 8th International Conference on Behavioral and Social Computing (BESC). pp. 1–5. IEEE (2021). https://doi.org/10.1109/BESC53957.2021.9635271
55. Natali, C., et al.: Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by design through frictional decision support systems. In: CEUR WORKSHOP PROCEEDINGS. vol. 3481, pp. 15–19 (2023)
56. Norman, D.: The design of everyday things. Doubleday Business (1990)
57. Ohm, P., Frankle, J.: Desirable inefficiency. Fla. L. Rev. **70**, 777 (2018)
58. Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: An attentional integration. Human factors **52**(3), 381–410 (2010). https://doi.org/10.1177/0018720810376055
59. Park, J.S., Barber, R., Kirlik, A., Karahalios, K.: A slow algorithm improves users' assessments of the algorithm's accuracy. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–15 (2019). https://doi.org/10.1145/3359204
60. Pierce, J.: Undesigning interaction. Interactions **21**(4), 36–39 (2014). https://doi.org/10.1145/2626373
61. Pierce, J.: In tension with progression: Grasping the frictional tendencies of speculative, critical, and other alternative designs. In: Proceedings of the 2021

CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2021). https://doi.org/10.1145/3411764.3445406
62. Rastogi, C., Zhang, Y., Wei, D., Varshney, K.R., Dhurandhar, A., Tomsett, R.: Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW1), 1–22 (2022). https://doi.org/10.1145/3512930
63. Riva, P., Aureli, N., Silvestrini, F.: Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? Acta Psychologica **229**, 103681 (2022). https://doi.org/10.1016/j.actpsy.2022.103681
64. Sambasivan, N., Veeraraghavan, R.: The deskilling of domain expertise in AI development. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2022). https://doi.org/10.1145/3491102.3517578
65. Sayre, J.W., Toklu, H.Z., Ye, F., Mazza, J., Yale, S.: Case reports, case series–from clinical practice to evidence-based medicine in graduate medical education. Cureus **9**(8) (2017). https://doi.org/10.7759/cureus.1546
66. Sengers, P., Boehner, K., David, S., Kaye, J.: Reflective design. In: Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility. pp. 49–58 (2005). https://doi.org/10.1145/1094562.1094569
67. Serdar, C.C., Cihan, M., Yücel, D., Serdar, M.A.: Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia medica **31**(1), 27–53 (2021). https://doi.org/10.11613/BM.2021.010502
68. Sharma, M., Sharma, C.: A review on diverse applications of case-based reasoning. Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019 pp. 511–517 (2020)
69. Shiraishi, J., Li, Q., Appelbaum, D., Doi, K.: Computer-aided diagnosis and artificial intelligence in clinical imaging. In: Seminars in nuclear medicine. vol. 41, pp. 449–462. Elsevier (2011). https://doi.org/10.1053/j.semnuclmed.2011.06.004
70. Tenner, E.: The efficiency paradox: what Big Data can't do. Vintage (2019)
71. Vaccaro, A.R., Oner, C., Kepler, C.K., Dvorak, M., Schnake, K., Bellabarba, C., Reinhold, M., Aarabi, B., Kandziora, F., Chapman, J., et al.: Aospine thoracolumbar spine injury classification system: fracture description, neurological status, and key modifiers. Spine **38**(23), 2028–2037 (2013). https://doi.org/10.1097/BRS.0b013e3182a8a381
72. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–15 (2019). https://doi.org/10.1145/3290605.3300831
73. Wickens, C.D., Clegg, B.A., Vieane, A.Z., Sebok, A.L.: Complacency and automation bias in the use of imperfect automation. Human factors **57**(5), 728–739 (2015). https://doi.org/10.1177/0018720815581940
74. Wilbanks, J.: Design issues in e-consent. Journal of Law, Medicine & Ethics **46**(1), 110–118 (2018). https://doi.org/10.1177/1073110518766025
75. Wolfe, C.R., Britt, M.A.: The locus of the myside bias in written argumentation. Thinking & reasoning **14**(1), 1–27 (2008). https://doi.org/10.1080/13546780701527674