The need to move away from agential-AI: empirical investigations, useful concepts and open issues

Federico Cabitza[a], Andrea Campagner[a], Carla Simone[b]

[a] *University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy*
[b] *University of Siegen, Siegen, Germany*

Abstract

We propose a novel approach to human interaction with artificial intelligence systems (HAII), alternative to the mainstream dyadic one where humans and AI are seen as interacting agents. Through two quantitative experiments and two qualitative in-field case studies, we show that the mainstream HAII paradigm presents potentially harmful design shortcomings as it can trigger negative dynamics such as automation bias and prejudices. Our proposal, on the other hand, is grounded in the Computer-Supported Cooperative Work literature, in which AI can be conceived as a component of a Knowledge Artifact (KA). This consists of an ecosystem of knowledge creation tools whose goal is to support a Ba (after Nonaka), i.e. a collective of competent decision makers. We highlight the cooperative nature of decision making and the AI functionalities that a KA should embed. These include eXplainable AI solutions, aimed at facilitating appropriation, but also functionalities that enable reasoning in a collaborative setting. Finally, we discuss how moving intelligence and agency from individual agents to the human collective can help to mitigate the shortcomings of dyadic HAII (e.g., deskilling), re-distribute responsibility in critical tasks, and revisit the HAII research agenda to align it with the needs of increasingly wide, heterogeneous and complex teams.

Keywords: Machine Learning, Intelligent Systems, Artificial Intelligence, Knowledge Artifact, Ba

1. Introduction

The current main paradigm of human interaction with Artificial Intelligent Systems (HAII) stems from within the classical man-machine paradigm. This conceives individual human beings as interacting with AI systems, and makes human-machine dyads [56, 58, 69] a primary concept. In this ambit, two main narratives clash and complement each other. On the one hand, the AI system is regarded as a *tool* that empowers individuals, by augmenting their cognitive abilities in decision making tasks [55]. On the other, the AI system is also seen as an *autonomous agent* that can replace the human actor in tedious, repetitive, critical, dangerous and error-prone tasks [3, 32].

These two narratives are the extremes on a full range of possible configurations, where humans and AI systems are considered as agents that can even collaborate as teammates [29, 42]. However, all the stances stem from a common cognitivist stream of research [36]. From this standpoint, it is plausible to regard AI systems as (more or less) autonomous agents, which are capable of expressing legitimate (but obviously not infallible) interpretations and classifications of given instances and cases [1, 58].

The reliability and accuracy of these agential interpretations (often presented in terms of advice or recommendations) are thus pursued by virtue of their potential to improve human cognition beyond some of its known shortcomings. Examples of the latter include the many unconscious biases observed in decision tasks in psychology research, such as, priming effects, framing effects, availability biases, and false causality [32, 45, 55, 57]. On the other hand, the cognitivist approach also studies the *appropriation* of AI systems by their prospective users, together with the related problems due to a lack of trust, transparency or fairness [56]. These latter aspects, in particular, have been emphasized within research related to eXplainable [55, 57, 70] (XAI) and human-in-the-loop AI [26, 29], as well as co-evolving socio-technical systems [25, 36].

However, also these latter approaches have not focused much on the potentially harmful effects of AI support on the cognition of human individuals who interact with these systems [33, 59, 70]. These effects include *automation bias* [22] (i.e., the propensity of humans to favor suggestions from AI systems); *automation complacency* [4] (i.e., the unjustified satisfaction with an AI system, possibly leading to non-vigilance, or poor detection of malfunctions); and "prejudices against the machine" [11, 74] (i.e., the opposite of automation bias, the propensity to reject the suggestion from AI systems).

As argued by Shneiderman [59], these latter issues may be intrinsic to the dyadic HAII paradigm due to its inability to properly take into account the relational, collaborative and often implicit shared knowledge of typical decision-making tasks into which AI systems are deployed as support. In this sense, the dyadic paradigm and its solutions to achieve appropriation could be understood as a design flaw, because they detract from other important design aspects. In fact, Shneiderman [59] observed that HAII should be centered on *Humans in the group; computers in the loop*, rather than *humans in the loop.* A similar point has also been made by Thomas Malone, when he suggests that "we should move away from thinking about putting humans in the loop to putting computers in the group" [40]. The above mentioned problems may also be potentially harmful from the point of view of the human-AI fit, because they may trigger negative dynamics such as complacency, deskilling and avoidance of responsibility [8, 17, 23, 66].

In this article, we thus propose an alternative approach to HAII, encompassing the above mentioned works in XAI and co-evolving systems and inspired by the Computer-supported Cooperative Work (CSCW) literature as well as Shneideman's understanding of AI systems as "supertools and active appliances, rather than teammates, partners, and collaborators" [59]. In short, we conceptualize AI systems as *Knowledge Artifacts* [16] (KA), that is composite tools aimed at supporting *knowledge work* [14] within a collective of competent users. The users are expected to share some common ground and cooperative goals, to improve their competencies (learn), and to produce knowledge in their collaborative tasks: What Nonaka [46] refers to with the Japanese term Ba[1].

The paper explores the above mentioned shortcomings of the dyadic HAII paradigm as well as the conceptual framework of our proposed approach and its implications for the HAII field. The rest of this article will be structured as follows. In Section 2 we discuss the related work on HAII and CSCW. In Section3 we focus on the limitations and potentially harmful consequences of the traditional HAII paradigm, through two quantitative experiments (in Section 3.1) and two qualitative in-field case studies (in Section 3.2). In Section 4, we detail our proposal and its implications for the HAII field. Finally, in Section 5, we summarize our findings and delineate possible lines of future research.

2. Background, Motivations and Related Works

Human work, irrespective of whether it is performed individually or not, involves rich networks of relationships between humans, spanning through time and space. These networks of relationships build the unique *social context* that gives each activity its meaning in terms of interdependence, shared resources and information, and social relationships [54].

The design of any technology aimed at supporting human work, therefore has to account for this context and for the practices that human beings establish to have things done in their setting. This means recognizing the fact that actors are *competent* and that technology should not disrupt the consolidated practices where these skills are expressed, but rather preserve and corroborate them. This is the main tenet that is common to various research lines within the CSCW literature, from participatory design [67] and ethnomethodology [51], to the recent idea of *practice-based computing* [50].

Studies in this field recognize that the introduction of technologies in any work setting influences its existing practices, and vice versa, practices can influence the way in which the technology is adopted and used. The successful adoption of a technology thus depends on the extent this mutual influence results in a balanced process that is usually called *appropriation* [20]. This process goes beyond mere adoption and use of a specific technology, as it includes learning new uses [52], possibly even beyond the designers' intentions.

Appropriation is thus a delicate process [19, 61]. It is facilitated if the technology is easy to use. But this is not enough. The technology must also fit the real needs of, and show clear advantages to, the users. If appropriation is, as it should be, a goal for a technology, then the design process has to be guided by this goal. In this sense, early user involvement through various forms

[1] Literally, the term Ba (場) denotes either a space or place; an occasion or situation; a field (both as intended in *physics*, and as intended in the phrase *field of knowledge*); or also the area where cards are laid out (in card games).

of *participatory design* (e.g., co-design [63] or continuous design [31]), makes appropriation more likely to happen as users play an active role, so that they can express their needs and provide their suggestions when the design process is still ongoing, and at the same time they can anticipate the the embedding of the technology into their situated practices.

The above mentioned observations regarding appropriation have been widely illustrated in situations where the issues at stake concern knowledge-intensive activities such as learning and decision making, i.e. *knowledge work* [2]. Research in CSCW questions the interpretation of knowledge as a set of decontextualized objects that can be exchanged and shared, and of memory as a mere repository of these objects [72]. Consequently, it also challenges all AI approaches that take up or support this interpretation. In fact, knowledge is (the result of) a social process, including the learning and self-training that occurs in technology appropriation. Knowledge work can produce data, as a sort of persistent trace of its activities (e.g., the formalized output of consensually agreed decisions), and only these traces can be digitally represented, and stored in some material format, with a necessarily limited account of the context in which they were produced.

The increasingly wider adoption of complex technologies, such as those embedding ML models and components, makes their appropriation even more problematic since ML models are often *black boxes* [68] whose output is difficult to interpret in the context in which they are used and, often uncritically, trusted. Since any technology incorporates representations of various kinds (e.g., to describe and classify the entities at stake), and "no representation is complete and permanent [. . . as it] is the snapshot of historical processes in which different viewpoints, local contingencies and multiple interests have been temporarily reconciled" [21], a limited accessibility of the complex representations ML uses to produce its output makes its appropriation more problematic.

A recent trend in the HAII literature has thus been to design AI systems that make their internal representations more explicit, so as to make appropriation easier. This trend includes the current research related to XAI [57] and human-in-the-loop AI [26, 29], as well as the field of human-AI co-evolution [25, 36]. These techniques can be effective at improving the interaction between human users and AI systems in certain contexts. Holzinger et al. [26, 29] described the effectiveness of human-in-the-loop approaches on a case study based on the Traveling Salesman problem, as well as the use of graph neural network for explainable and causable AI. Shin [57, 58] studied the effects of explainability and causability on AI appropriation. Zhang et al. [75] showed the positive effects of explainable AI techniques for the appropriation of AI systems. Lewis et al. [36] studied the effects of co-evolution on the appropriation of AI systems into coaching practice. Similarly, Navidi et al. [44] described a co-evolution method based on reinforcement learning which showed promising results in simple problems.

Nonetheless, recent research [37, 59, 70] has highlighted how also such methods can lead to problematic appropriation, primarily due to the emergence of biases that may affect the interaction between the human users and the AI systems and that can be reinforced by XAI methods [33], e.g. automation bias and the “white box paradox” [6].

Thus the open question we wish to address is: how can we exploit the computational power of ML-based intelligent systems while still guaranteeing their appropriation within a group of practitioners to support their decision making and knowledge-based activities? In the next section, we will report about some experimental studies that shed light on the need for this move and how to prepare for it. We present the methods and main results of two user studies conducted in an experimental and controlled environment. We also report the main findings from two ecological case studies, which are real-world projects in which intelligent systems had been deployed and validated by various teams of practitioners.

3. The quantitative and qualitative case studies

This section details two independent groups of case studies, whose goal is to highlight the above mentioned limits of the traditional HAII paradigms. We combine and present both quantitative and qualitative findings to make a sounder point, that we analyse in the following sections.

The first group of case studies presents two quantitative experiments that highlight how the traditional approach to HAII leads to biases, such as prejudice against the machine or automation bias

The findings of these experimental studies are then complemented by two qualitative case studies reporting the main findings from two projects, where ML-based tools were introduced into a large multinational IT company. Observations obtained through discussions with one key member of the two project teams [60] are discussed also in light of the two quantitative experiments, so as to highlight the importance of properly considering the practices and contexts in which AI systems are to be adopted.

3.1. The quantitative user studies

3.1.1. First study: prejudice against the machine in ECG reading

In the first experiment 73 clinicians, with varying proficiency in reading ECGs, were asked to express their opinions on three clinical cases which had been selected from the ECG Wave-Maven[2] database by two board-certified senior cardiologists. The answers were collected through an online multi-page questionnaire developed on LimeSurvey[3] (version 3.21). Each page of the questionnaire reported a brief clinical description and a single ECG as well as the form containing questions.

At the beginning of the session, each participant was told the following backstory: “The hospital where you are (fictionally) employed has recently acquired an advanced and certified AI decision support system associated with a validated diagnostic accuracy of approximately 96–97%. A colleague of yours, a

[2] https://ecg.bidmc.harvard.edu/maven/mavenmain.asp

[3] https://www.limesurvey.org/

senior cardiologist with 26 years of experience in ECG reading, has disagreed with the interpretation supplied by the AI system in regard to three cases and has asked for you and others to be *second opinion readers*" [11].

The platform set out the diagnoses proposed by both the human colleague and the AI, and asked the participants which one they wanted to confirm, or if they disagreed with both. Respondents were also requested to assess the plausibility of each diagnosis (on a semantic differential rating scale, from 1 'very low' to 4 'very high'). However, the two diagnoses were both equally wrong and randomly assigned to either the AI or the human colleague. Moreover, in order to avoid order bias, half of the sample read the diagnosis proposed by the AI before the diagnosis by the human colleague, while the other half read the diagnoses in the opposite order. In the last page of the online questionnaire, the respondents were asked their gender, main specialty, work experience in that specialty, and self-assessed proficiency in ECG reading on a 3-level rating scale (from 1, basic skills, to 3 advanced skills).

Our goal was to assess how biases such as *conformity bias* or *automation bias* could affect human decisions in case of erroneous advice. In order to evaluate if the first diagnosis really influenced the readers, or if they simply confirmed the readers' own interpretations, after a due wash-out period of four months we also administered a second survey, where the same panel of experts were supposed to consider the same three ECGs (in different order) but without receiving any prior diagnosis, neither by AI nor a human colleague. To assess the presence of any statistically significant difference, we applied the Mann-Whitney U test, a non-parametric procedure to compare ordinal groups of measurements; and the χ^2 test, to compare sample proportions. In all cases, significance was determined based on a 95% confidence level (i.e. $\alpha = .05$).

Results. We collected 246 unique interpretations by 75 ECG readers. With regard to the perceived plausibility of the diagnosis given by either the human colleague or the AI system, the respondents who discarded these options considered them significantly less plausible than those who trusted them by confirming one of them (Mann Whitney test, $U = 10684$, p-value < 0.001). The respondents who discarded the diagnoses did not found them to be significantly different in terms of plausibility (Mann Whitney test, $U = 1999$, p-value $=$ 0.68). Conversely, the respondents who confirmed one of the diagnoses found the diagnosis given by the human colleague significantly *more plausible* than that given by the AI (Mann Whitney test, $U = 8863$, p-value $= 0.002$). More in general, the whole group of respondents deemed the diagnoses proposed by the AI system significantly *less reliable* than those proposed by the human colleague (Mann Whitney test, $U = 20149$, p-value $= 0.011$). These results are reported in Figure 1 in terms of a box plot which highlights how the average perceived reliability of the human expert was significantly higher than that of the AI, and almost higher than the third quartile of the latter distribution of responses.

Gender and proficiency in ECG reading were found to affect the perceived accuracy of AI: female readers considered the AI diagnoses significantly *more reliable* than male readers (Mann Whitney test, $U = 3299$, p-value $= .045$). The

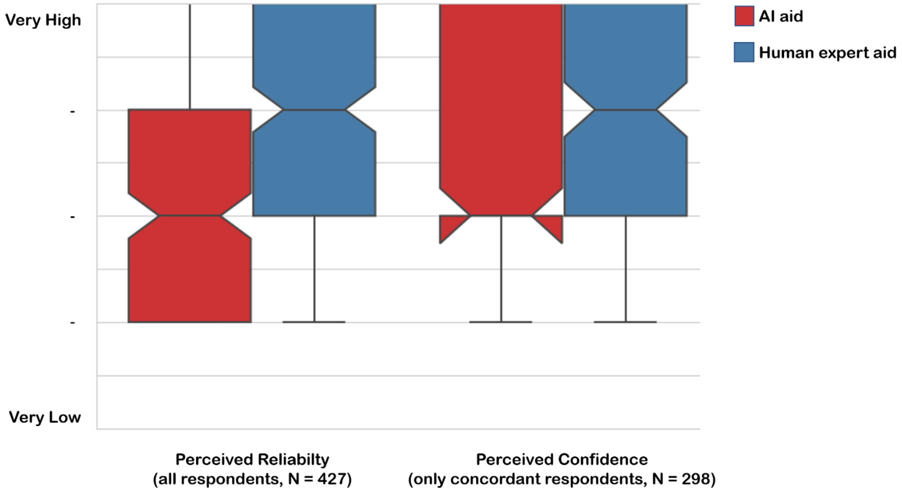

Figure 1: Box plots of the perceived reliability of the diagnoses given to the respondents, for all respondents (on the left), and for the respondents who agreed with one of the proposed diagnoses (on the right). Responses were given on an ordinal 4-value scale from 'very low' to 'very high'. Notches indicate the 95% confidence intervals of the medians (no overlap indicates statistically significant difference). N indicates the number of plausibility values recorded, not the number of respondents who gave those values.

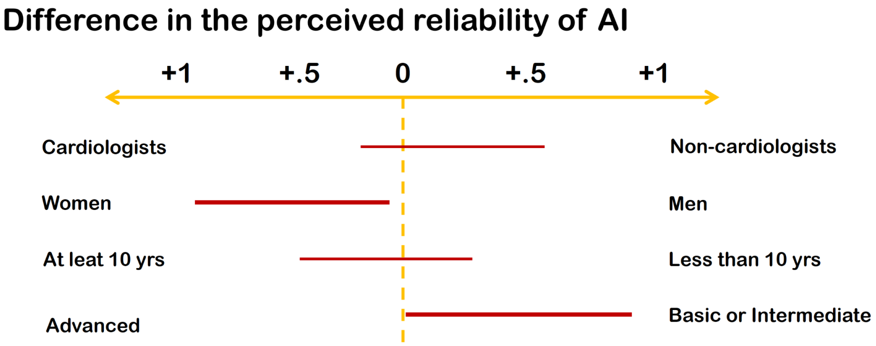

Figure 2: Confidence intervals of the difference of the average perceptions of reliability of the AI advice. If an interval does not contain the vertical line denoted as 0, the corresponding means are significantly different. The interval is represented closer to the category of respondents who attached the higher reliability to the AI advice. Thicker lines are associated with statistical significance.

same holds for the less skilled readers compared to the more skilled ones (Mann Whitney test, $U = 2213$, p-value $= .048$). No other profile characteristic was found to have a significant influence on the perceived reliability of AI.

The confirmation rate was high. The respondents agreed with one of the two (wrong) diagnoses most of the time (70%). Cardiologists and more experienced readers discarded the diagnoses significantly more often than other respondents (35% vs 18%, p-value = .016; and 50% vs 25%, p-value = .005, respectively). No other profile characteristic was found to have a significant influence on the confirmation rate. Although the human and the machine's diagnoses were both wrong, the participants chose the human diagnosis significantly *more often* than the AI one (57% vs 43%, p-value = .044).

As a consequence of the results, we can observe that human clinicians tended to trust *agential AI* systems less than their human colleagues, thus highlighting the presence of a "prejudice against the machine" [11]: a phenomenon that has also been observed in other settings (e.g., [74]). As shown in Figure 2, this prejudice was found to be stronger in the cardiologists and in the more experienced ECG readers, than in other medical specialists. The same figure also highlights how both female (as compared to male) and less skilled (as compared to highly skilled) readers significantly over-estimated the AI reliability. Such potential users may thus tend to over-rely on the AI advice (as they may deem the AI system to be more accurate than they are), ultimately leading to both automation bias and long-term deskilling [17]. Trust in *agential AI* systems may thus depend on the characteristics not so much of the trustee (that is, the *agential AI*) but rather of the trustors (i.e., the users of the AI system). Consequently, appropriation of these systems by their prospective users can depend not so much on the performance of the AI but on the degree to which their users properly understand and trust such AI systems [57].

3.1.2. Second study: when conversation improves group performance more than computational aids

In the second experiment, we involved six small teams of human participants, with three to five members each. The teams were involved in a gamified experiment - a geography trivia test - whose goal was to compare three human-AI cooperative work protocols: one in which the AI acted as an agent, a second one in which the AI only acted as a trigger for discussion among team members, and a third protocol that did not encompass any AI support. Each protocol was applied to the trivia test of two teams.

The participants were selected from a group of 200 students of a HCI class by means of a preliminary questionnaire that assessed general proficiency in geography-related questions. The preliminary questionnaire was implemented and administered through LimeSurvey and had a total of 18 questions, of various levels of difficulty. To discourage cheating, the participants had a time limit of 14 seconds, for each question.

From the whole group of 200 students we selected a sub-sample of 16 participants, who were grouped into three types of teams, based on their performance in the preliminary questionnaire:

1. Two groups (called the Centaurs) of three people each, selected from the respondents who scored best in the preliminary questionnaire (average group accuracy ∼70%).
2. Two groups (the Aurigae) of five people each, selected from respondents with average-to-low score in the preliminary questionnaire (∼45%).

3. Two groups (the Legionaries) of five people each, selected from respondents with an even lower score ($\sim 40\%$).

Each group was involved in a moderated session, set up on Google Meet. Each group had to answer 24 geography trivia questions, with only medium-to-high difficulty questions. The participants in each group, for each question, had to first provide their answer independently on a Google Form questionnaire with four options (the correct answer was one of the four options). After giving their individual answers, the groups had to produce a definitive group answer, which was recorded by the authors before passing to the next question. To this end, the three group types were supported differently and asked to cooperate through a different collaboration protocol.

1. The Centaurs were told they would receive the aid of an "agential" AI that was supposed to be as accurate as the best respondent among them (85%). A single participant (rotating) was tasked to answer each question: this participant had to give their answer and could then ask for the support of the AI or not; in cases of disagreement in regard to the right answer to give, this participant could ask for the help of another teammate, through a structured discussion protocol similar to a second-opinion setting;
2. The two Aurigae groups were also assigned an AI support, but with a lower accuracy (46%). The advice of the AI system was given to the groups before they could collaboratively discuss (within a limited time frame) the available answer options;
3. The two Legionaries groups were not assigned any AI, but were allowed to discuss each question freely.

For each group and each geography trivia item, we recorded both the answers provided by each respondent, independently of the other members of the same group, as well as the official answer given after a consensus had been reached. Both the individual and collective responses were used to assess the performance of the groups in terms of accuracy, i.e. proportion of correct answers.

The main goal of this second experiment was to assess how AI support, especially in case of erroneous advice, and collaboration could impact on group decisions and accuracy; and whether different modes of cooperation and AI support could have different effects on the performances of the groups.

The statistical significance of the findings was assessed either through the Kruskall-Wallis H test (when simultaneously comparing the three groups) or the χ^2 test (when comparing two proportions). Correction for multiple hypothesis testing, in order to avoid false discoveries, was performed through the Dunn-Bonferroni procedure. In all cases, statistical significance was determined based on a 95% confidence level (i.e., $\alpha = .05$).

Results. The results of the experiment, in terms of the accuracy of the individual participants, the average group accuracy (averaging the former accuracy scores), and the collective group accuracy are reported in Tables 1, 2, 3 (for the Centaurs, Aurigae and Legionaries groups, respectively) 4 and in Figure 3.

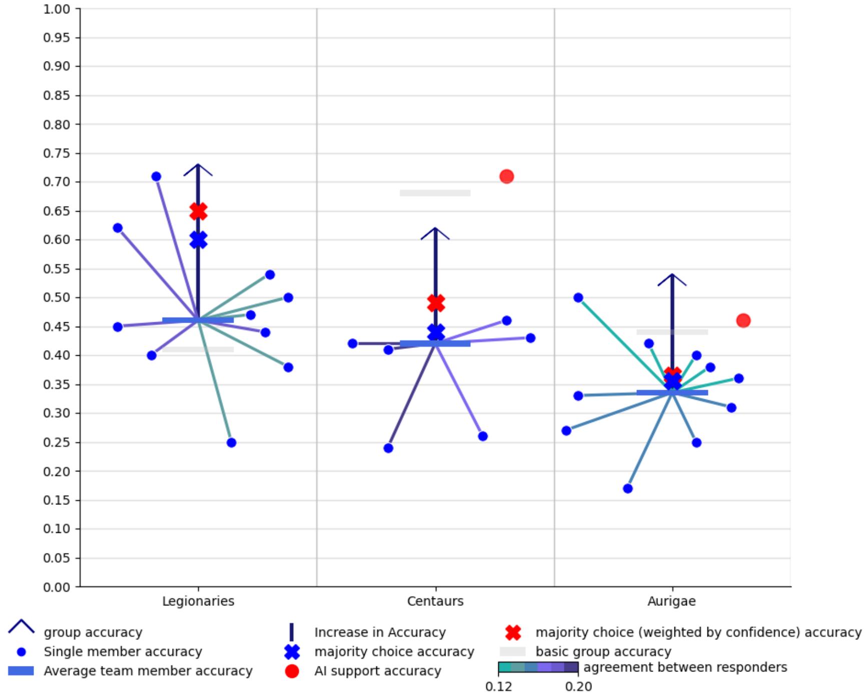

Figure 3: Results of the geography trivia experiment. The performance of the three groups (Legionaries, Centaurs and Aurigae) are reported on the left, the center and the right, respectively; while the y axis refers to accuracy. In each of the three graphs, the blue dots represent the accuracy of the single group members; the grey bar represents the average accuracy of the group members in the preliminary assessment; the blue bar represents the average group accuracy; while, the arrow pointer represents the group accuracy (i.e. the accuracy of the answers reported after collective discussion). Then, the vertical line connecting the blue bar to the arrow pointer represents the increase in accuracy due to discussion. The accuracy of the AI support for the three groups is represented as a red dot.

Table 1: Average results, for the three group types in regard to the single responses; along with the group accuracy scores in regard to their consensus responses.

Groups	pre-test (Avg.)	pre-discus. (Avg.)	Group perf.
Centaurs	67.3%	37%	62.5%
Aurigaes	43.8%	34.3%	54%
Legionaries	40.9%	47.8%	73%

The performance of the individual group members in the preliminary evaluation was significantly different (Kruskal-Wallis test $H = 10.75$, p-value $= 0.001$), and rightly so, since we wanted Centaurs to be more accurate (in geography

Table 2: Results for the Centaurs groups

respondent ID	pre-test	pre-discus.
1A	72%	42%
2A	67%	42%
3A	61%	25%
Avg.	67%	36%
Group	-	58%
respondent ID	pre-test	pre-discus.
1B	82%	25%
2B	61%	46%
3B	61%	42%
Avg.	68%	38%
Group	-	67%

Table 3: Results for the Aurigae groups

respondent ID	pre-test	pre-discus.
1A	56%	29%
2A	47%	33%
3A	33%	29%
4A	17%	17%
5A	44%	29%
Avg.	39%	27%
Group	-	54%
respondent ID	pre-test	pre-discus.
1B	56%	38%
2B	44%	42%
3B	44%	38%
4B	47%	50%
5B	50%	38%
Avg.	48%	41%
Group	-	54%

trivia) than the members of the other groups. Likewise, the average performances of the three group types (pre-discussion accuracy) were significantly different. The differences in accuracy between Centaurs and both Aurigaes (p-value = 0.008) and Legionaries (p-value = 0.001) were significant. Conversely, the difference between Aurigaes and Legionaries was not significant (p-value = 1).

With respect to differences in performance between the pre-discussion accuracy scores and the performance in the preliminary evaluation, the difference was significant for both Centaurs and Aurigaes (p-value equal to 0.002 and 0.014, respectively), but not significant for the Legionaries (p-value = 0.120). In particular, the performance of both Centaurs and Aurigaes worsened in the

Table 4: Results for the Legionaries groups

respondent ID	pre-test	pre-discus.
1A	44%	42%
2A	35%	72%
3A	56%	46%
4A	28%	63%
5A	39%	42%
Avg.	40%	58%
Group	-	88%

respondent ID	pre-test	pre-discus.
1B	50%	38%
2B	50%	50%
3B	39%	46%
4B	39%	54%
5B	29%	25%
Avg.	41%	43%
Group	-	58%

pre-discussion assessment. On the other hand, even though the performance of the Legionaries improved, this improvement was not statistically significant.

For the collective group performance, the statistic of the Kruskal-Wallis test was $H = 6.829$, with a p-value $= 0.033$. As the null hypothesis was rejected we also performed the post-hoc Dunn-Bonferroni test. The differences between Centaurs and both Aurigae (p-value $= 1$) and Legionaries (p-value $= 0.522$) were not significant. On the other hand, the difference between Aurigae and Legionaries was significant (p-value $= 0.026$).

With regard to the differences between the collective (post-discussion) group performance and the average individual performance, the difference for the Centaurs (p-value $= 0.026$) and Legionaries groups (p-value $= 0.028$) was significant, while the difference for Aurigae was not significant (p-value $= 0.08$).

Our results highlight that the groups where the AI system acted as a trigger for discussion (i.e., the Aurigae) reported a greater improvement than the groups where the AI was assigned an agential role, discussion was curbed, and decision-making was performed by human-AI dyads (i.e., the Centaurs). Indeed, for the Centaurs, we observed a decrease in performance compared to the baseline one reported in the first individual assessment. These findings suggests that AI systems may be more useful when they are not intended as oracles [70] but rather as triggers for collective discussion, in which humans have to rely on their transactive memories [47] and social skills. This could also be due to the mitigation of the potential biases that the use of an *agential AI* can induce [6, 74] (as we observed in the first experiment) and of their potential detrimental effects on the performance of the team members. As further evidence toward this point, the increase in performance was even greater for the groups where AI was not present at all (i.e., the Legionaries). This suggests that collaboration and discussion in themselves were more important, in terms of their beneficial effect on the performance of the groups, than the availability of any type of AI support, due to its potential role in priming or biasing the teammates' opinion.

3.2. The qualitative case studies

3.2.1. First case: A difficult appropriation

In the first project [7] a sales analytics tool was developed for an internal group of sellers of a global cloud IT infrastructure-as-a-service (IaaS) to help them prioritize sales opportunities, reduce churn and defections, and target particular cloud-service offerings to expand this new service. In the first phase of the ML development, a single predictive feature was identified as a good predictor of client accounts that were likely to defect in the next six-month period. Although this simple algorithm yielded meaningful results, the project team wanted to explore more advanced machine learning methods to improve the precision and accuracy of the predictions. The developed ML model had a precision higher than 90%, while its accuracy was above 80%.

Despite these promising results, the validation in the field of the ML-based analytics highlighted several important lessons. The initial *risk of defection* model followed a simple formula which was easy to interpret and convey: the sellers could assess the reasonableness of the outcomes by inspecting the data and posit a reason why certain accounts were on the risk-of-defection list. On the other hand, the more advanced ML model took into account multiple features and their non-linear relationships. As a result, the model was difficult to interpret, as it was difficult for the sellers to see the direct link between the data and the predictions, and impossible for the designers to explain in plain language how the model arrived at those predictions. Thus, despite the efforts made to improve accuracy and precision, the usefulness of the analytics was not clear to the sellers, who eventually stopped using the ML tool.

Although the predictions were associated with high probability scores, these scores by themselves were not sufficient to suggest what actions should be taken in response. In fact, there were potentially many factors outside the model that, in the current practices of the sellers, were influencing the prediction. In addition, sometimes the sellers could not understand why a client account had been removed from the list for each kind of risk. What possible external actions had been taken in the meanwhile, and by whom, to improve the account status? In order to get this kind of contextual information from the sellers, the interface was endowed with specific questions to help designers tune the model and, in the future, have it provide recommendations about useful steps that the sellers might take to correct the situation. Regrettably, since there were no existing work practices for how this information could be recorded, the sellers rarely provided such feedback.

The project team's conclusion was that "data analytics are often portrayed as offering ready-to-use solutions for those with the data and the expertise to put them to work. But our recent experience has humbled us and exposed us to a myriad of challenges, even obstacles, that must be overcome to realize the full potential of enterprise analytics".

3.2.2. Second case: The need for a flexible interaction with the tool

The same company started another project [71] to help a pool of professionals (called *architects*) to find the optimal IT configuration in response to a complex document, called Request for Service (RfS), provided by their clients.

The resulting system used ML techniques to classify RfS contents, so as to extract the configuration requirements and to look for the optimally satisfying IT configuration and setting. Although the architects were highly motivated and active in proposing improvements to make the system successful, its validation highlighted some serious shortcomings.

The ML-based classification of the RfS contents was not capable to account for the often implicit relationships between the requirements, nor for other conditions that did not directly impact the technical solution but, nevertheless, should be part of the response to the client. Moreover, the classification was not able to recognize repetitions in the content and classified those repetitions either erroneously or inconsistently. To this matter, the project team made the following observation: "Outliers, patterns, variance — given that algorithms process data like math – how can these forms of computation be leveraged by people as they collaborate with algorithmic systems? This is not a matter of push-button automation, but instead a more concerted, reciprocal relationship with the tool's ML capabilities".

Furthermore, since design work practices were closely coordinated through frequent conference call meetings, the architects complained that the system did not consider that they usually shared tips and tweaks with each other, thereby calibrating their solution designs as they move closer to a firm solution. "Mutual and collaborative intelligibility is the overarching goal in these practices rather than a focus on completeness [53]".

An additional issue regarded the variability and incremental definition of the client requirements and the difficulty in exploiting the model's outputs in this recurrent situation. The system was designed to be able to keep track of changes but did not provide any documentation about the motivations as to why those changes had occurred. The architects' positive attitude towards the system led them to suggest improvements on how the RfS contents classification could be presented to improve their ability to parse the RfS document, by including explicit information about the levels of uncertainty and their motivations; and on how to open the models toward a truly collaborative environment, where additional contextual information is made available and ready to improve the usefulness of the response. In this project the final conclusion was that in "the development of ML models in industrial settings where only small datasets exist, a paradigm of ongoing refinement and iterative feedback is necessary, whereas one-shot, highly complex, trained models are not feasible".

Our interviews with the project team highlighted how the architects experienced a continuous conflict between the challenge to creatively figure out a way to appropriate the technology, possibly by modifying their current practices, and the difficulty in bringing the technology to an adequate level of effectiveness in their work context [65] and that this was one of the reasons why the project came to an end.

In conclusion, what these two projects highlight is the need to conceive ML-based solutions so as to take into account their target setting. Work contexts pose different requirements for a successful appropriation than contexts where existing practices play a less crucial role. In addition, the continuous interaction with, and involvement of, prospective users is a key element to understanding how the potential of an ML model can be adapted to the different practices in different work domains and how to improve the appropriation of this new, powerful and complex technology.

4. A novel integrative proposal and opportunities

4.1. The concepts of Knowledge Artifact and Ba

The main findings from our quantitative studies (see Table 5 for a schematic summary), confirm the literature on the biases implied by agential AI in decision support [11, 22, 33, 74] (case 1) and also suggest that group discussion can leverage collective intelligence and lead to a better performance than super-human AI (case 2). The two qualitative case studies, on the other hand, shed light on the need to contextualize (in the broadest sense) the AI advice, so as to support the users' creative appropriation of the technology.

In fact, the two qualitative case studies highlight that the current shortcomings and potential risks of ML-based AI support are due to the lack of attention, in the design phase, to the tacit and collective dimensions of the knowledge work in which these tools must be embedded and integrated [59]. This underlies our proposal to recast "intelligent" support as a collection of complementary and different *functions*, that support a community of competent practitioners: what Shneiderman has called "supertools and active appliances "[59]. In this subsection, we describe the two core elements of our proposal: the design-oriented construct of the Knowledge Artifact and the Ba.

A Knowledge Artifact [15, 16] is defined as being any computational artifact that is *collaboratively created, maintained, used and/or purposefully adopted* to support knowledge-oriented social processes (among which knowledge creation and exploitation, collaborative problem solving and decision making). Thus, the intended aim of a KA is to support action in cooperative settings according to its *negotiated structure* and *contingent content*, as well as allow for an *affordable, continuous and user-driven maintenance and evolution*, of both its structure and content, at an *appropriate level of underspecification*.

Let us focus first on user-driven evolution and underspecification, as they are constitutive of any KA. User-driven evolution can be related to the continuous and incremental evolution of the classification model and its training dataset, as well as to specific protocols that combine users and the learning platform, as highlighted in the literature on co-evolving systems [36]. The resulting classification models thus represent a joint effort across two communities of users: i) the raters involved in the ground truthing and annotation of the first training set; and ii) the users that have to appropriate the classification model into their decision practices. In this respect, approaches inspired by *active learning*, *online learning*, or *conformal prediction*, could combine the efforts of the annotating raters and the end users, to facilitate the appropriation of the ML models.

More in detail, in the context of a KA, *active learning* approaches could be useful to identify significant cases in the work environment of the end users, and update the embedded ML models to better capture the salient characteristics of these cases. Thus, these approaches can be exploited to perform a *domain adaptation*. This entails transferring the capabilities of the ML models (as developed in the context of the annotating raters) to the domain of interest of the end users, and empower these users by asking them to pro-actively contribute toward the development and effective deployment of the ML models.

Similarly, *online learning* and related [44] approaches may allow the community of end users of the classification model to *progressively adjust* the capabilities of (and the representations embedded within) the ML model, so as to make the latter more aligned with the relevant domain context, and also to successfully realize a continuous, incremental improvement of this model.

Finally, *conformal prediction*-based protocols [5] can enable a *robust* appropriation, insofar as they allow the ML models to resort to underspecified (i.e., partial) classifications when sufficient information is not yet available.

This property of conformal prediction protocols is related to the second peculiar aspect of a KA, i.e., *underspecification*. Underspecification has important and far-reaching implications in regard to how ML models should be deployed in cooperative settings. In fact, *underspecification* entails not considering ML models as (possibly fallible) sources of oracular advice and classification truths. Instead of clear-cut classes, the underspecified output of a classification system should regard the following: multiple probability scores (in multi-class cases); the capability to abstain when these scores are lower than a predefined (also user-driven) threshold [5, 34, 12]; and, most importantly, a set of different indications about the proposed classification. These indications should be redundant and possibly contradictory, incomplete and yet complementary, as if they were pieces of a jigsaw puzzle which it is the task of the human interpreter to understand and reassemble, even if many parts are missing or in the wrong place. All these pieces of information are part and parcel of a KA, what in [18] and [10] was defined as *knowledge evoking information* or *inspired inefficiency*. It is through this output that a KA stimulates knowledgeable actions, it prompts human reasoning about the solution of a specific issue, or it facilitates recalling pertinent and effective solutions, or known facts, rules, past cases, and events. All these can, almost recursively, become new parts of the KA, if appropriately documented and shared among its users.

These observations also makes it clear that a KA cannot be decoupled from the tacit knowledge that exists within a community of knowledgeable practitioners: their knowledgeable behaviors with regard to new cases are facilitated and promoted by KAs that draw on the documented activities of that community, or of close and similar communities (such as that of the raters who labelled a *ground truth* training set). Thus, the concept of KA naturally complements the concept of Ba we mentioned in Section 1.

In Japanese, Ba denotes a place or field. It should not be intended as "a physical space" but rather "where [and when] information is interpreted to become knowledge [. . .] a space nexus that simultaneously includes space and time" [46], i.e. a socio-technical configuration where the key concept is interaction. However, what differentiates a Ba from the physical or virtual place where ordinary human interaction occurs is knowledge creation. A Ba is a group of interacting people that internalize, socialize, externalize knowledge together, learn and develop competent behaviors when confronted with unseen cases or unsolved problems. A KA is then any concrete support for this community of competent actors in terms of affordances that facilitate interaction, and information that facilitates knowledge circulation and ratification. Our point is that a KA can encompass machine learning capabilities, as well as XAI and human-in-the-loop functionalities [27], once these are not aimed at exhibiting intelligent agential behavior, but rather at informing decision makers and helping them combine the ML models' output with other information sources that have been created to account for the local context, language, idiosyncratic attitudes and perceptions, and professional approach, as we will see in Section 4.2.

4.2. From intelligent agents to functions for human knowledge

We are not the first ones to point out the risks of *agential AI* in human assemblies [9, 17, 66, 74]. In this paper, however, our aim is to clarify the concept of agential AI as the main source of harm, a point also raised by Shneiderman and Wiens [59, 70]. This latter point is connected, as we have argued so far, to dysfunctional and innate (i.e., not easily eradicable) reactions in human actors within a collaborative setting [33, 74].

Therefore, if the nature of knowledge as a dynamic process and social practice must be recognized and leveraged, a more collective perspective should be adopted with respect to traditional stances where both humans and intelligent systems are seen as agents, i.e. entities that act and interact [69]. We thus believe that computational tools should be designed to facilitate cooperative agency, as an emerging property of human collectives. We call these collective ensembles Ba, to avoid the static, structural meaning that is strongly associated with other terms such as community, team or group. As previously discussed, a Ba is a working team, whose members argue and discuss cases, facilitated by computational tools whose main aim is to help them externalize, socialize and produce knowledge, also by getting access to the results of past good practices and effective decisions, i.e. computational tools we have denoted as KAs. In this light, we see the output given by intelligent systems that embed some ML model as a sort of *knowledge evoking information* [18], and multiple models (and output typologies) as corresponding multiple functions of a KA.

To illustrate the richness of functions that ML-driven systems could add to a KA used within a Ba, we describe a set of such functions. Most of them can be traced back to the classification model, or augment it by means of functions inspired by the XAI literature [27], but others refer to ancillary models that operate on the same training set to complement the output of the main ML model [13]. We then can envision a KA that:

1. Gives indications about the dataset used to train the classification model, which could be useful to inform the users about the reliability of the AI systems and its training set. Such indications include the amount of missing data, the number of the raters involved, their expertise, the ground truthing protocols, or the agreement among the raters [35];
2. Highlights the areas of an image (as in saliency or activation maps [62]), or the attributes of a record [12] that are more informative. In line with eXplainable and human-in-the-loop AI, these pieces of information could help identifying the most relevant "patterns" for the case in hand, or could be used for debugging and auditing AI systems [70];
3. Selects the most similar and most different cases from the training set, and presents them with respect to the new instance to make a decision about [12, 30]. These functions could aid analogical and contrastive reasoning [43], by pointing out relevant similarities and differences between the case in hand and past knowledge;
4. Allows the users to change some of the attributes of a given case to simulate small modifications in the representable characteristics. Such modifications could be implemented by embedding a causal model [49] describing the relationships among the variables and features of interest. In doing so, users could be supported in performing counterfactual reasoning [28]; in reasoning about interventions that can change the observed phenomenon; and in reasoning about causability [27, 57];
5. Embeds a model that has learnt the perceived average difficulty of the training instances, as perceived by the raters involved in the annotation process. This could be used to estimate the difficulty of a new case in light of the previous ones, so as to inform the users about potential risks of misclassification or draw attention to cases that need more attention;
6. Embeds a model that has learnt the perceived average confidence of the raters in their provided annotations. This model could be used to estimate how certain would the raters be when confronted with a new case;
7. Embeds a model that has learnt the agreement scores, measured with some appropriate metrics (such as Krippendorff's Alpha [35]), on the training cases. This model could then be used to estimate the agreement on any new instance. The output of the three models outlined in points 5, 6 and 7 could be used to "simulate" how the group of raters involved in the annotation process, i.e. a team of knowledgeable colleagues and representatives of the same community of practice, would express their consensus interpretation in a new case. For instance, a pattern such as "high difficulty, low confidence and low agreement" should warn the decision makers that the ground truth annotations for such a case could be considered as dubious and, as such, the classification model could get the case wrong and potentially mislead them;
8. Embeds a computational model that supports the argumentation within the Ba [39]. This function would help in the automated extraction of arguments from textual annotations [38], in the enumeration of the arguments (including the information proposed by the ML-embedding KA)

that have been made related to the classification of particular cases, and in reasoning about the network of relationships among these arguments (e.g. which arguments contradict, or support, each other; which groups of arguments can be considered coherent) [64]. Such a model could facilitate the discussion among the users, which we showed in the geography trivia experiment to be beneficial for the appropriation of AI systems.

5. Conclusion

> *The Analytical Engine has no pretensions whatever to originate any thing. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths. Its province is to assist us in making available what we are already acquainted with. This it is calculated to effect primarily and chiefly of course, through its executive faculties; but it is likely to exert an indirect and reciprocal influence on science itself in another manner. For, in so distributing and combining the truths and the formula of analysis, that they may become most easily and rapidly amenable to the mechanical combinations of the engine, the relations and the nature of many subjects in that science are necessarily thrown into new lights, and more profoundly investigated. This is a decidedly indirect, and a somewhat speculative, consequence of such an invention.*
>
> —Ada Lovelace, *1843, From Note G, p. 722*

In this paper, we presented an alternative way to think of and use AI support, inspired by the literature on XAI [27] and co-evolving systems [36] and centered on the CSCW perspective. Our approach recognizes the cooperative nature of work and the illusory nature of knowledge not only as an information asset, but also as an individual achievement, in favor of an idea of knowledge as social practice. We have discussed how this approach could help avoid the main decision biases that are usually associated with intelligent support. We have shown how at the same time it would facilitate teamwork, as it entails collaboration and a cooperative use of AI rather than its individual use.

We have thus adopted concepts that have already been proposed in the HCI and CSCW literature, but which to the best of our knowledge, have never been combined together: the KA and the Ba, which denote a computational tool that enables knowledge-oriented collaboration, and the collaborative ensemble itself, respectively. Our aim is to relate them to the use of intelligent systems in collaborative decision making. We see KAs as coherent sets of heterogeneous functions that are supportive of knowledge-related processes and facilitate the creation and development of a Ba, i.e. a group of competent practitioners engaged in interpretative tasks who need to get access to some past experiences of their community of practice to make sense of new cases.

Interpreting AI tools as KA functions within (and for) a Ba allows us to overcome two important limitations of the agential approach. This latter approach conceives AI systems as (more or less) autonomous entities that can make mistakes or, conversely, do things right according to how accurate their judgment is. However, this means that the design and development of such AI systems will tend to favor small improvements in this abstract accuracy metric rather than get a broader socio-technical perspective [25]. On the other hand, AI systems that are proposed as intelligent entities can, once their interactive capabilities have been deemed to be sufficiently articulated, be also considered as *teammates* [42] of individual users, who are then considered as partners engaged in abstract and unrealistic human-machine dyads.

To these two intertwined ideas we counterpose the intuition of Ada Lovelace, reported in the epigraph above. This should not be seen in foretelling terms, but rather admonitory ones. In analogy with Lovelace's statement about the Analytical Engine, AI systems should be conceived as computational processes that support us by "making available what we are already acquainted with" but would not be available otherwise, through convenient interactive tools. This shift entails evaluating the overall task where humans make their decisions (possibly using AI tools and their advice), not the abstract AI performance alone. In turn, this also entails considering the effectiveness of decision making, also accounting for the consequences of the actions taken (or not taken); its efficiency, above all in terms of timeliness and of the resources involved; as well as the satisfaction of the decision makers, which can be assessed by interviewing them or with ad-hoc questionnaires administered after each decision [37]. Following Lovelace's idea of going against the idea of AI as a partner that is capable "of anticipating any truth", our approach looks at human action as emerging from a network of interdependencies and relations of a collaborative nature. This network is situated in a socio-technical environment, where the information provided by the AI (or by multiple AI systems) is only one of the possible elements considered in a collective deliberation. This entails seeing AI as a trigger in this collaborative consultation, and as an enabler or facilitator of team-mating and co-reasoning, and therefore to optimize its design for this role. Indeed, as observed by Heaven "An AI system needs to fit into a process where sources of uncertainty are discussed rather than simply rejected" [24]. AI tools should then be seen as memex-like tools [73] that facilitate access to past decisions and are active parts of a hybrid transactive memory [13, 47]: a similar idea was also provocatively proposed in [48], when the authors speak of AI as a tool for knowledge extraction, or what they call a *Nooscope*.

In conclusion, while the approach we propose can be seen as a sort of egg of Columbus to mitigate the risk that decisions that can impact a person's life are distorted by label bias at the level of the ML model [70], or by some cognitive bias at the level of the individual decision maker [32], we believe that further research should be aimed at investigating if other kinds of biases, which are more typical of collective arrangements [41], such as 'groupthink', 'social loafing', 'group polarization' and 'escalation of commitment' would impact AI-supported Bas, especially in delicate sectors such medicine, law, access to credit, customer relationships and human resource selection. Further research is necessary in this line of research, as advocated by this paper.

List of Abbreviations

- AI - Artificial Intelligence
- CSCW - Computer Supported Cooperative Work
- HAII - Human Interaction with Artificial Intelligence
- HCI - Human-Computer Interaction
- KA - Knowledge Artifact
- ML - Machine Learning
- XAI - Explainable Artificial Intelligence

References

[1] Michael D Abràmoff, Danny Tobey, and Danton S Char. Lessons learned about autonomous ai: finding a safe, efficacious, and ethical path through the development process. *American journal of ophthalmology*, 214:134–142, 2020.

[2] Mark S Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. Sharing knowledge and expertise: The cscw view of knowledge management. *Computer Supported Cooperative Work (CSCW)*, 22(4-6):531–573, 2013.

[3] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. In AI we trust? perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3):611–623, 2020.

[4] J Elin Bahner, Anke-Dorothea Hüper, and Dietrich Manzey. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, 66(9):688–699, 2008.

[5] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[7] Jeanette Blomberg, Aly Megahed, and Ray Strong. Acting on analytics: Accuracy, precision, interpretation, and performativity. In *Ethnographic Praxis in Industry Conference Proceedings*, volume 2018, pages 281–300. Wiley Online Library, 2018.

[8] Raymond R Bond, Tomas Novotny, Irena Andrsova, et al. Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of electrocardiology*, 51(6):S6–S11, 2018.

[9] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017.

[10] John Buschman. The efficiency paradox: What big data can't do. *Journal of Information Ethics*, 29(2):107–111, 2020.

[11] Federico Cabitza. Biases affecting human decision making in ai-supported second opinion settings. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 283–294. Springer, 2019.

[12] Federico Cabitza, Andrea Campagner, Davide Ciucci, and Andrea Seveso. Programmed inefficiencies in dss-supported human decision making. In *Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, September 4–6, 2019, Proceedings*, volume 11676, page 201. Springer, 2019.

[13] Federico Cabitza, Andrea Campagner, and Edoardo Datteri. To err is (only) human. reflections on how to move from accuracy to trust for medical ai. In *ITAIS 2020: Proceedings of the XVII Conference of the Italian Chapter of AIS Organizing in a digitized world: Diversity, Equality and Inclusion, Pescara, Italy*, 2020.

[14] Federico Cabitza, Andrea Cerroni, and Carla Simone. The knowledge-stream model. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-Volume 3*, pages 367–374, 2014.

[15] Federico Cabitza, Gianluca Colombo, and Carla Simone. Leveraging underspecification in knowledge artifacts to foster collaborative activities in professional communities. *International Journal of Human-Computer Studies*, 71(1):24–45, 2013.

[16] Federico Cabitza and Angela Locoro. Made with knowledge: Disentangling the it knowledge artifact by a qualitative literature review. In *KMIS 2014 - Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 64–75. INSTICC, 2014.

[17] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.

[18] Federico Cabitza and Carla Simone. Affording mechanisms: an integrated view of coordination and knowledge management. *Computer Supported Cooperative Work (CSCW)*, 21(2-3):227–260, 2012.

[19] Deborah S Debono, David Greenfield, Joanne F Travaglia, et al. Nurses' workarounds in acute healthcare settings: a scoping review. *BMC health services research*, 13(1):1–16, 2013.

[20] Alan Dix. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21*, pages 1–4, 2007.

[21] Elihu M Gerson and Susan Leigh Star. Analyzing due process in the workplace. *ACM Transactions on Information Systems (TOIS)*, 4(3):257–270, 1986.

[22] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

[23] Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

[24] WD Heaven. Google's medical AI was super accurate in a lab. Real life was a different story. *MIT Technology Review*, 2020.

[25] Robert Holton and Ross Boyd. 'Where are the people? What are they doing? Why are they doing it?' Situating artificial intelligence within a socio-technical framework. *Journal of Sociology*, page 1440783319873046, 2019.

[26] Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.

[27] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[28] Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. Towards multimodal causability with graph neural networks enabling information fusion for explainable ai. *Information Fusion*, 71:28–37, 2021.

[29] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49(7):2401–2414, 2019.

[30] Eyke Hüllermeier. Towards analogy-based explanations in machine learning. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 205–217. Springer, 2020.

[31] Suhas Govind Joshi and Tone Bratteteig. Assembling fragments into continuous design: On participatory design with old people. In *Scandinavian Conference on Information Systems*, pages 13–29. Springer, 2015.

[32] Daniel Kahneman, AM Rosenfield, L Gandhi, and T Blaser. Noise. *Harvard Bus Rev*, pages 38–46, 2016.

[33] Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, page 103458, 2021.

[34] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):1–6, 2021.

[35] Klaus Krippendorff. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112, 2011.

[36] Lise Lewis and David Clutterbuck. Co-evolution: exploring synergies between artificial intelligence (AI) and the supervisor. In *Coaching Supervision*, pages 200–216. Routledge, 2019.

[37] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[38] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25, 2016.

[39] Luca Longo and Lucy Hederman. Argumentation theory for decision support in healthcare: A comparison with machine learning. In *International conference on brain and health informatics*, pages 168–180. Springer, 2013.

[40] Thomas W Malone. How human-computer'superminds' are redefining the future of work. *MIT Sloan management review*, 59(4):34–41, 2018.

[41] Russell Mannion and Carl Thompson. Systematic biases in group decision-making: implications for patient safety. *International Journal for Quality in Health Care*, 26(6):606–612, 2014.

[42] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Christopher Myers. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors*, 60(2):262–273, 2018.

[43] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[44] Neda Navidi and Rene Landry. New approach in human-ai interaction by reinforcement-imitation learning. *Applied Sciences*, 11(7):3068, 2021.

[45] Ben R Newell and David R Shanks. Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, 37(1):1–19, 2014.

[46] Ikujiro Nonaka and Noboru Konno. The concept of "ba": Building a foundation for knowledge creation. *California management review*, 40(3):40–54, 1998.

[47] Edward T Palazzolo. Transactive memory and organizational knowledge. *Communication and organizational knowledge: Contemporary issues for theory and practice*, pages 113–132, 2010.

[48] Matteo Pasquinelli and Vladan Joler. The nooscope manifested: Artificial intelligence as instrument of knowledge extractivism. *Artificial Intelligence and Society*, 2020.

[49] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[50] Dave Randall, Markus Rohde, Kjeld Schmidt, and Volker Wulf. Introduction: Socio-informatics—practice makes perfect? In *Socio-Informatics*. Oxford University Press, 2018.

[51] David Randall. Investigation and design. *Socio-Informatics: a practice-based perspective on the design and use of IT artifacts (1st ed). Oxford University Press, Oxford, UK*, pages 221–241, 2018.

[52] Mike Robinson and Liam Bannon. Questioning representations. In *Proceedings of the Second European Conference on Computer-Supported Cooperative Work ECSCW'91*, pages 219–233. Springer, 1991.

[53] Kari Rönkkö. " yes-what does that mean?" understanding distributed requirements handling. In *Social Thinking-Software Practice*, pages 223–241, 2002.

[54] Kjeld Schmidt. Riding a tiger, or computer supported cooperative work. In *Proceedings of the Second European Conference on Computer-Supported Cooperative Work ECSCW'91*, pages 1–16. Springer, 1991.

[55] Procheta Sen and Debasis Ganguly. Towards socially responsible ai: Cognitive bias-aware multi-objective learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2685–2692, 2020.

[56] Donghee Shin. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior*, 109:106344, 2020.

[57] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

[58] Donghee Shin. Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *Journal of Information Science*, 2021.

[59] Ben Shneiderman. Human-centered AI. *Issues in Science and Technology*, 37(2):56–61, 2021.

[60] Carla Simone, Wagner Ina, Müller Claudia, Wiebert Anne, and Wulf Volker. *Future-proofing: Making Practice-based IT Design Sustainable.* Oxford University Press, 2021. In press.

[61] Carla Simone, Angela Locoro, and Federico Cabitza. Drift of a corporate social media: The design and outcomes of a longitudinal study. In *Organizing for the Digital World*, pages 189–201. Springer, 2019.

[62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[63] Marc Steen. Co-design as a process of joint inquiry and imagination. *Design Issues*, 29(2):16–28, 2013.

[64] Hannes Strass and Adam Wyner. On automated defeasible reasoning with controlled natural language and argumentation. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 765–773. AAAI Press, 2017.

[65] Christine T. Wolf and Jeanette L. Blomberg. Ambitions and ambivalences in participatory design: Lessons from a smart workplace project. In *Proceedings of the 16th Participatory Design Conference 2020-Participation (s) Otherwise-Volume 1*, pages 193–202, 2020.

[66] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, pages 1–16, 2021.

[67] Ina Wagner. Critical reflections on participation in design. In *Socio-Informatics*. Oxford University Press, 2018.

[68] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364, 2019.

[69] Marcus Westberg, Amber Zelvelder, and Amro Najjar. A historical perspective on cognitive science and its influence on XAI research. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 205–219. Springer, 2019.

[70] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[71] Christine Wolf and Jeanette Blomberg. Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019.

[72] Chi-Lan Yang, Chien Wen Yuan, and Hao-Chuan Wang. When knowledge network is social network: Understanding collaborative knowledge transfer in workplace. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[73] Richard Yeo. Before memex: Robert Hooke, John Locke, and Vannevar Bush on External Memory. *Science in Context*, 20(1):3, 2007.

[74] Ryosuke Yokoi, Yoko Eguchi, Takanori Fujita, and Kazuya Nakayachi. Artificial intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. *International Journal of Human–Computer Interaction*, pages 1–10, 2020.

[75] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Table 5: Tabular summary of the studies described in this article. CAWI stands for Computer-Assisted Web self-Interview, DSS stands for Decision Support System, MRI stands for Magnetic Resonance Imaging, ECG stands for Electrocardiogram.

Study ID	Application Domain	Experimental Design	Method	Investigated Effect	Main findings
1	ECG reading (diagnosis, reporting)	independent measures between-groups	CAWI w/ 75 cardiologists	automation bias; prejudice against the machine	trust/distrust, automation bias correlated w/ expertise, gender and AI familiarity
2	collaborative Geography trivia	independent measures between-groups	mixed: psychometric questionnaire; observation	performance improvement	more discussion-oriented, less AI supported protocols associated w/ higher improvement and satisfaction
3	corporate Business Intelligence DSS	qualitative post-mortem analysis	unstructured interviews w/ key informants	reasons for low use, low satisfaction	lack of interpretability, disregard of collaborative practices; need to support followup action
4	corporate Business Intelligence DSS	qualitative post-mortem analysis	unstructured interviews w/ key informants	reasons for low use, low satisfaction	Need to integrate context w/ advice; Promote/support creative use of the technology; disregard of collaborative practices