April 2022

Open, multiple, adjunct. Decision support at the time of relational AI

Federico CABITZA [a,1] and Chiara NATALI [b]
[a] *University of Milano-Bicocca, IRCCS Istituto Ortopedico Galeazzi*
[b] *University of Milan, University Vita-Salute San Raffaele*

Abstract.
In this paper, we consider some key characteristics that relational AI should exhibit to enable decision hybrid agencies that include subject-matter experts and their AI-enabled decision aids, especially when these latter ones have been developed by following a machine learning approach. We will hint at the design requirements of guaranteeing that AI tools are: *open*, *multiple*, *continuous*, *cautious*, *vague*, *analogical* and, most importantly, *adjunct* with respect to decision making practices. We will argue that especially *adjunction* is an important condition to design for. *Adjunction* entails the design and evaluation of *human-AI interaction protocols* aimed at improving AI usability, that is decision effectiveness and efficiency, while also guaranteeing user satisfaction and human and social sustainability, as well as mitigating the risk of automation bias, technology over-reliance and user deskilling. These high-level aims are compatible with the tenets of a relational approach to the design of AI tools to support decision making and collaborative practices.

Keywords. Relational Artificial Intelligence, Decision support, Machine Learning, Interaction protocols, Usability

1. Introduction

Nearly 25 years ago, Giorgio De Michelis [1] wrote a little book in Italian, entitled "Aperto, molteplice, continuo: gli artefatti alla fine del Novecento" (Open, multiple, continuous: artifacts at the end of the twentieth century), where he dealt with the design of artifacts, in particular digital artifacts, and the perception, and use of them. Twenty years later, we believe these aesthetic categories should be taken again in regard to the design of digital decision support, or better yet, for the design of the interaction ways in which we (designers) have these computational tools, which often embed very complicated, almost inscrutable correlative models developed by means of Machine Learning (ML) methods to map their input and output together, offer their advice to situated decision makers. We will make a point that in order to both exhibit artificial intelligence (a seeming autonomy

[1] Corresponding author: federico.cabitza@unimib.it:
This is the extended version of the article published in the proceedings of the first conference on Hybrid Human Artificial Intelligence (HHAI'22) Amsterdam, the Netherlands, 13-17 June, 2022. To cite this article, please use the following format: Cabitza, F., & Natali, C. (2022). Open, multiple, adjunct. Decision support at the time of relational AI. In HHAI2022: Augmenting Human Intellect (pp. 243-245). IOS Press. https://doi.org/10.3233/FAIA220204

in produce effective behaviors in front of unexpected situations) and promote augmented intelligence (in decision makers facing the very same unexpected situations), ML-based decision support systems must be: open, multiple, continuous, cautious, vague, analogical and adjunct.

We mean that an ML-based decision support must be:

1. **open**, that is open to the external environment, or open-loop: in particular, the system should be capable to update its reference data (i.e., the ground truth upon which its models have been trained) and, consequently, its correlative models, so as to cope with ever-changing environment and mitigate the risk of errors due to concept drift [2]. This entails a tighter relationship with users, which is not just unidirectional – the machine that gives humans advice – but rather it is bidirectional, in that the user provides feedback on the correctness and usefulness of the recommendations and the relevance and sensibility of any explanations, with the machine that updates or recalibrates its logic accordingly.
2. **multiple**, in that the system should not propose to users single pieces of advice or limited its output to clear-cut categories but rather multiple and complementary indications (e.g., by proposing classes and the associated confidence scores), or even possibly identical and diverging pieces of advice by different competing models (e.g., models optimized for sensitivity, specificity, discriminative performance or utility [3]), what we denoted as Janic AI[4] to hint at the potential for ambiguity or opposition to unveil a more comprehensive (and perhaps unattainable) truth. This requirement of multiplicity, instead of being aimed at confusing users, should reflect the intrinsic complexity and ambiguity of the phenomenon upon which to support them, and mitigate phenomena such as automation bias [5], that is over-reliance on their advice, or the fallacious appeals to "algority" (or *algorithmic authority*). Multiplicity is also related to the output of the divergent phases [6] of computer-aided generative design [7,8].
3. **continuous**, in the sense hinted at by De Michelis[1], for whom it is continuous what "connects two phenomena, two different contexts without interruption" (p. 43) and regards "any situation in which openness and/or multiplicity has been developed." (p.45). Thus, instead of giving just discrete pieces of information, either numerical values or labels associated with new instances and cases, almost in an oracular manner, a continuous decision support system should allow for the dialectical interaction between it and its users in order to allow for the exploration of the causal factors, possible explanations and effects on their output, which derive from a full range of small differences in the digital representation of those instances and cases. Thus, continuity is a requirement related to connecting different representations of the same piece of advice (see e.g., [9,10]) and also to the capability to enable counterfactual or contrastive reasoning [11], by allowing small perturbations in the input so as to provide a full range of pieces of advice reflecting those modifications and hence to give hints about the causal connections between input representations and the machine's predictions.
4. **cautious**, in that the system should express a judgment only when its confidence about its output is sufficiently high, or above a threshold that depends on task criticality, the risk of failure, or users' expertise or preferences. When confidence is lower than this threshold, the model should rather abstain [12] and express its temporary inadequacy in supporting the user for the case at hand.

5. **vague**, in that, like in the case of multiplicity, the system should not limit itself in providing one best option, but rather promote reflection in expert users by proposing multiple pertinent classes for the case at hand, guaranteeing high confidence in that the list or interval of values given contains the right answer, like in *conformal prediction* settings [13,14].
6. **analogical**, in that the system should foster analogical thinking in experts, by presenting to them the most (or the least) similar cases to the case at hand, according to their correlative models and some similarity metric [15], and by inviting the users to reflecting what similarity (or dissimilarity) could suggest the correct answer on the basis of the original true labels associated with those cases [16]. This can also entail systems showing what features of the new case at hand are "normal" (that is similar to most of the available past cases), and hence expectable [17], or the other way round, abnormal, irrespective of the final label associated with the case. We call this kind of AI *Epimethean* (i.e., reasoning turned backward, to the past) in opposition to traditional predictive computing, which instead can be called Promethean (because aimed towards the future, facing forward) in its being a support for *prospection*: an Epimethean AI considers the experience upon which the AI has been trained a sort of reified knowledge of the past, and the classifier model produced with machine learning techniques, a compressed archive of knowledgeable decisions to tap in to inform future decisions (past presents informing the present futures, instead of the present futures informing the future presents[18]).
7. **reflective**, in that the system should foster reflection, e.g., by not limiting itself to providing full-fledged answers or advice, but also challenging users about their confidence (e.g., "how sure are you from 1 to 10?") and asking pertinent (according to the case) questions that promote counter-factual reasoning (such as, "would you change your mind if the patient were 10 years younger?") or the pursuing of alternative options (such as, "what would it be your second best option?"), as discussed in [19] where the concept of "reflection machine" was proposed as a way to increase meaningful human control over decision support systems, and also as a way to avoid fixation [17].
8. **adjunct**, a concept that we will discuss in more detail in the next section. Here we anticipate that, while the above mentioned features are intrinsic to and related to specific functionalities and affordances of the decision support, being adjuct is a matter of how the tool is integrated in the main decision process and the surrounding work practices.

2. For a theory of AI adjunction

Let us start from giving the definition of *Human - AI Collaboration Protocol* (HAI-CP): this is an integrated set of rules and policies that stipulate the use by competent practitioners of AI-exhibiting tools to perform a certain task or do a certain job; for instance, a HAI-CP can determine what data are made available to the AI tool; what data the AI is supposed to provide users with; at what step in an articulated process, and in what order with respect to the work of human beings.

In short, HAI-CPs stipulate how human decision makers should interact with the machines that support them. The requirement of *adjunction* regards HAI-CPs in which

the AI component is relegated to the "edges of the process", that is it is put on ancillary tasks and off the critical path or under continuous human oversight. Thus, a theory of adjunction invites to focus on the process-oriented and relational aspects of the joint action of humans and machines working together. This entails, among other things, the evaluation of *human-plus-machine* systems as a whole, and therefore avoiding isolating the performance of one (kind of) component from the performance of the other one (like in traditional validation and testing of ML systems).

The uneven effect exerted on output quality by computational power on the one hand, and process quality on the other, is illustrated by the *Kasparov's law* [20]. According to the chess grandmaster Garry Kasparov, *"A clever process beats superior knowledge and superior technology"* [21]. Or, more formally put:

1. Weak Human + Machine + Better Process > Strong Machine;
2. Weak Human + Machine + Better Process > Strong Human + Machine + Inferior Process

However, we should consider that AI systems can also undermine the decision-making skills they are supposed to improve, by inadvertently inducing forms of complacency, deskilling, and avoidance of responsibility. [22,23,24,25,26]

Moreover, over-reliance on technology and mental outsourcing [27] has been shown to permanently affect cognition [28], and remodel the brain as an instance of negative neuroplasticity that has been likened to digital dementia [27]. Our term of choice for the pathological dependency on decision support is *decision atrophy* [4]. Following Thomas W. Malone's insight, we should ask ourselves how people and computers could be connected as to operate more intelligently as a group than any individual, team or machine before them. [29]

The answer we propose is the *adjunction* design approach, as a mild form of *exclusionary action* that is close to what Pierce calls displacement [30]: in adjunction, human-AI interaction protocols are conceived to purposefully move the AI support to the background or to a role of "second opinion" giver [31], after that an official (and registered) decision has been already made by single human decision makers or by small teams of decision makers.

This concept stands in a critical relation (but not opposition) with that of human oversight [32], which however deals with problems only after they occur [33].

2.1. Overcoming Dyadic HAII

Human oversight is the first of the seven pillars of trustworthy AI delineated by the *European Union Digital Strategy's Ethics Guidelines for Trustworthy AI* [34]. It entails technical human control over a deployed AI system, as well as accountability for the whole development and deployment process, which is based on human judgement [35]. Oversight aims to ensure that AI systems do not threaten personal autonomy or have other negative consequences; to do so, it requires the adoption of suitable human-machine interfaces to assist humans in supervising high-risk AI systems as they operate. [36,37] In the *IEEE Ethical Aligned Design (Version 2)* report, oversight is linked to the notions of transparency and accountability [38], while the European Commission AI Watch attaches to oversight the keywords "human control" and "human in the loop" [39]. The *Ethics Guidelines for Trustworthy AI* drafted in 2019 by the High-level Expert Group on Artificial Intelligence [40] pinpoint three possible governance mechanisms to ensure proper

oversight: the aforementioned human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approaches, with HITL generating the most interest in academia and beyond.[2]

Human-on-the-loop refers to the capability for human monitoring and intervention during a system's design cycle. The Human-in-command approach is centered on the user's capacity to supervise the AI system's activities and overall impact, while keeping control over whether or how the system is deployed in any given scenario. Finally, Human-in-the-loop is a method of human oversight where the user is allowed and even encouraged to intervene in every decision cycle of the system, even adjusting the algorithm or learning system itself [41]. This is not feasible, nor desirable, in many circumstances. [36] Since "human-in-the-loop" hybrid intelligence [42] can process highly unstructured information [43], it achieves performances that neither a human nor a computer could achieve on their own [44]. When the human confidence of the system output is low, for instance, humans can intervene and directly manipulate and improve the input, so to result in a feedback loop that increase the overall system's performance. The focus is on addressing challenges by strategically allocating tasks across AI-based machines and human agents [43].

"Human-in-the-Loop" typically implies the idea that computers will perform practically everything, but humans should be within reach in case some unexpected event happens. However, when workers' primary function is reduced to that of alienated auditors of the technology well-functioning, we become complacent and inattentive to the circumstances, hindering the situational awareness that the adjunction model aims at fostering (a phenomenon called automation complacency [45]).

The key difference between the adjunction model and the human-in-the-loop approach lies in the latter being an example of dyadic HAII paradigm [46], where humans and AI are seen as symmetrical interacting agents [22], and the latter one is not conceived as just a tool used by humans. However, the relational, cooperative, and even tacit collective knowledge of typical decision-making tasks into which AI systems are deployed as aids is not fully captured by dyadic HAII models. [22]

A theory of AI adjunction emphasizes the importance of AI-assisted humans taking ultimate responsibility for their decisions [47], as well as it recognizes both the cooperative nature of decision-making [48] and the distributed nature of cognition. The following section will describe the role adjunction can play in ensuring accountability. We will advocate for an AI-Decentered Humanity to be reached by cognitive friction and programmed inefficiency, as opposed to the Human-Centered AI advocated by proponents of the HITL, HOTL, and HIC methods.

2.2. Cognitive Friction and AI-Decentered Humanity

Human-centered AI, when aimed at smoothing out every instance of friction from our course of action, harbors the risk of engendering a gradual yet unavoidable degradation of the human attributes we value most in decision making: autonomy, intuition, and accountability[49,4].

[2] A simple search on Google Scholar can show this. As of March 10th, 2022, searching for "Human-in-the-loop" yielded about 43,700 documents. With 1,300 results "Human-on-the-loop" is much less discussed, while "Human-in-command" is close at approximately 1,000 results.

Thus, instead of a human-centered AI, we should try to achieve an *AI-decentered humanity* where AI is integrated into established work processes as an adjunct, rather than an essential, component [4]. If AI does have a detrimental influence on the attitude and learning processes of users, changing our minds, as users, is simpler than changing the AI itself (or demanding its vendors to change their plans). Raising awareness of the risks of automation is more straightforward than creating an ever more explicable, ethical or responsible AI, whatever this might mean. The adjunction theory, together with the concepts of programmed inefficiency and constructive distrust, provides a toolset for implementing this shift.

One example is anti-hedonistic machines [50]. They are devices used to increase friction between the user and a source of instant enjoyment, in the case such a gratification could have long-term detrimental implications [4]. The most straightforward examples of anti-hedonistic machines are cars equipped with brethalyzers, or Ignition Interlock Devices (IID), which prevent people from driving under the effects of alcohol; or, on a more playful note, morning alarm clocks that leap off one's nightstand, blaring and moving all around the bedroom to get the user out of bed [4]. In the words of Frischmann and Selinger (2018), *"some friction, some inefficiency, even some transaction costs may be necessary to sustain an underdetermined environment conducive to human flourishing."* (p. 141) [49]

While efficient AI aims to accelerate workflows and reduce relational friction, Adjunct AI can be given an opposing duty: slowing decision-makers down, making task fulfillment difficult or cumbersome, or even hindering people from performing a certain action [4].

This can be achieved via programmed inefficiencies [51], concrete modalities through which AI tools are steered towards a cognitive path where attention, vigilance and commitment are actively demanded by the technology (and protocol) to get things and work done. Programmed inefficiencies can be embedded in a system as to make the work process supported by the system purposefully less efficient, possibly longer, more difficult and less immediate than if performed without this feature. Other instances of this approach are deliberately engineered sources of friction [49], desirable inefficiencies [52], and inspired inefficiencies [53].

The main goal behind programmed inefficiencies is fostering constructive distrust [54] by arousing critical thinking, shattering the false impression of objectivity provided by algorithms[3], seeding questions about the outcome, nudging the user to look for more conclusive proof and fostering a sense of personal responsibility.

Thus, the suggestions provided by AI should be redundant and sometimes conflicting, partial and yet complementary. This knowledge-evoking information would act as jigsaw puzzle parts that the human interpreter must decipher and reconstruct [22]. Such cognitive-forcing functions, used during the decision-making process to disrupt heuristic thinking, would boost people's cognitive motivation for interacting analytically with the outputs, reducing overreliance on AI and improving performance [58].

In the AI adjunction framework, not only should AI present justifications or explanations for its recommendation, but this advice should also overcome the diffident questioning of the team members. The latter would be finally aware of the fallibility of the

[3]Following Sadin[55], we refer to this fallacy as the *alethic stance*: the inclination to regard the result of computational processes as more scientific, objective or neutral than any human output [56], despite significant counterevidence [57].

bot's advice, and this awareness has been shown to have a favorable effect on the entire team's performance [4,59]. The evidence collected by Christakis (2019) shows that teams that were assisted in virtual tasks by bots that were programmed to make sporadic mistakes, and admit them, consistently performed better than groups that relied on supposedly flawless bots. This was most likely because the former nudged the humans to improve communication and collaboration in the face of potential errors and uncertainty. [59].

2.3. Adjunction as Interaction Design

There are a plethora of techniques to promote individual, robust constructive distrust by design. In multi-class setting, always provide alternatives; in dichotomous settings, abstain if the confidence of the output is under a specific certainty threshold [12]; avoid direct advice (to avoid persuasion), but provide multiple counterfactual interpretations; embed two or more agonistic machine learning models [18], belonging to different families, trained on different representations, ground truths and parametrizations.

Other cognitive forcing functions are "checklists, diagnostic time-outs, or asking the person to explicitly rule out an alternative" [58], as well as not providing users with the AI suggestion by default, asking them to make an initial decision before confronting their insight with the output, and embedding longer waiting times[4].

A word of caution: although it is plausible to believe that increased friction would push users to think more critically and improve their performance, it could also reduce the perceived usability of the system, thus lowering its adoption [58]. Nonetheless, an adjunction theory of AI urges us to critically examine the corrosive motives of efficiency and comfort, and give priority to the efficacy and integrity of our knowledge work [4]. We need to find the ideal mix between cognitively engaging human-AI interaction protocols and the certainty that they will be adopted by teams of professionals. The degree of appropriation should be carefully assessed, not only on quantitative measures of performance (like error rates, throughput and execution times). We should also shed light on the use experience [61], trust and work environment arising from the feelings, attitudes and perceptions of the human colleagues towards the "Computer in the Group" [29].

Instead of requiring the evaluation of technology in isolation, as is the case today with the reporting of metrics based on machine error, the concept of adjunction leads us to consider the whole interaction protocol [22]. This entails the entire socio-technical system that adopts and deploys the AI, in terms of efficiency, efficacy, the satisfaction of both users and those affected (e.g. data subjects), sustainability (at the environmental, economic, social, human level) and cost-effectiveness.

Putting the computer in the group does not automatically imply that it should be considered a teammate [22]: in fact, Shneiderman referred to the "Teammate Fallacy", the belief that computers should be designed to function in teams because people do so [46].

From the lenses of adjunction, developers should be wary of infusing AI with simulated emotions [62,63,64] as the depiction of emotional responses such as regret for

[4]Previous studies have shown that slow algorithms actually help user accuracy[60]. Because of this, Buçinca et al. (2021) equate them to a cognitive forcing function. While waiting for the AI's suggestion, the user constructs their own hypothesis about the correct solution, and then evaluates the AI explanation to determine whether it supports their idea. [58]

mistakes elicited even higher levels of empathy and trust in the computer than ethopoeia did [65,66].

The adjunction perspective suggests that AI systems should be seen as "supertools and active appliances, rather than teammates, partners, and collaborators", as Shneiderman proposes [67]. Constructive distrust mitigates the harmful perception of AI output as "AI prophecies" [68], reducing it to one of many factors to be weighed in a group decision. This means viewing AI as a catalyst for collaborative discussion, and designing it accordingly [22].

3. Final remarks

In this piece we have surveyed some essential characteristics that AI systems should express when they are designed to support human decision making. These features are aimed at inspiring the design of human - AI interaction protocols through which a *humachine system* (or *humanchine* [69], i.e., a socio-technical system where humans and AI tools are tightly coupled) can exhibit some form of hybrid intelligence that is functional to some aim and sustainable in the long run. The reader could have noticed that we did not mention the concept of transparency, or interpretability (and the corresponding attributes: transparent, interpretable, explainable [70]) among the main categories that AI systems should reify, or at least inspire their design. The temptation to associate the characteristic of "being open" to the capability to be also *inspectable* by the human being (whether this be the user or anyone who has the right to ask for explanations regarding the functioning of the system if this latter has an impact on their rights or legitimate interests), could seem irresistible (also in light of the famous metaphor of "opening the black box" [16]). However, it would be so only to those who do not consider these instruments to be only tools, but rather agents that express legit judgements and interpretations [22] that, in order to have an effect, must be shared corroborated by detailed motivations or and sufficiently clear explanations. The history of Artificial Intelligence, ever since it appeared in the scientific debate through the reflections by Alan Turing on the subject [71], has always touched on the theme of persuasion and fiction[5]. However, we consider the explanations that such a system should produce to gain the characteristic of transparency or interpretability with respect to its output, simply put, an additional output [72]. This means that also explanations can be either right, and useful, or wrong, and useless or even harmful. Explanations can thus trigger reasoning useful to better understand the case in question [73], as well as misleading reasoning [72] or arguments that can make human beings more dependent, and thus trigger dangerous processes of deskilling [74]. These processes are the more harmful the more hidden by the apparent functionality to support the decision maker even more adequately: what we called the "white box paradox" [4].

An even more surprising absence might be the silence regarding the attribute that perhaps, most of all, indicates the feature that is widely and commonly considered necessary in next-generation decision support systems: trustworthiness. However, trust is one of the most relational attitudes that humans can feel and adopt when coping with others and, with some abuse of language, their tools. Right for its relational nature, trust cannot be embedded into any digital artifact but rather it *emerges* in use, depending on human

[5]Indeed, the Turing test essentially concerns the capability of an artificial system to deceive people on its own nature, or rather the ability of its creators to deceive those who interact with their creation.

attitudes and previous experience, and nurtured by reputation (of either the system or of those who produced them), and strengthened at each opportunity of use and interaction.

For this reason, since we want to take the relational nature of AI seriously and we believe that further and more important advances in the field of AI will not come from engineering and computer science, but rather from the scholarly disciplines focusing on interaction and collaboration design (such as cognitive ergonomics, human factors, human-computer interaction, computer-supported cooperative work) and various social theories (such as naturalistic decision theory, organizational theory, systems theory and cooperative game theory), in this contribution we presented some essential design-oriented concepts, and argued about their deeper significance for the design of effective, satisfactory and sustainable human-AI interaction.

References

[1] De Michelis G. Aperto, molteplice, continuo: gli artefatti alla fine del Novecento. Zanichelli, Milano; 1998.

[2] Zenisek J, Holzinger F, Affenzeller M. Machine learning based concept drift detection for predictive maintenance. Computers & Industrial Engineering. 2019;137:106031.

[3] Lu D, Tao C, Chen J, Li F, Guo F, Carin L. Reconsidering generative objectives for counterfactual reasoning. Advances in Neural Information Processing Systems. 2020;33:21539-53.

[4] Cabitza F. Cobra AI: Exploring Some Unintended Consequences;. .

[5] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association. 2012;19(1):121-7.

[6] Gabora L. Reframing convergent and divergent thought for the 21st century. arXiv preprint arXiv:181104512. 2018.

[7] Wu J, Quian X, Wang MY. Advances in generative design. Computer-Aided Design. 2019;116:102733.

[8] Chacón JC, Nimi HM, Kloss B, Kenta O. Towards the Development of AI Based Generative Design Tools and Applications. In: International Conference on Design, Learning, and Innovation. Springer; 2020. p. 63-73.

[9] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nature Medicine. 2020;26(8):1229-34.

[10] Rundo L, Pirrone R, Vitabile S, Sala E, Gambino O. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. Journal of biomedical informatics. 2020;108:103479.

[11] Verma S, Dickerson J, Hines K. Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:201010596. 2020.

[12] Campagner A, Cabitza F, Ciucci D. Three–way classification: Ambiguity and abstention in machine learning. In: International Joint Conference on Rough Sets. Springer; 2019. p. 280-94.

[13] Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. Springer Science & Business Media; 2005.

[14] Campagner A, Cabitza F, Berjano P, Ciucci D. Three-way decision and conformal prediction: Isomorphisms, differences and theoretical properties of cautious learning approaches. Information Sciences. 2021;579:347-67.

[15] Keane M. Analogical mechanisms. Artificial Intelligence Review. 1988;2(4):229-51.

[16] 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. 2020;4(1):1-7.

[17] Klein G. Snapshots of the Mind. MIT Press; 2022.

[18] Hildebrandt M. Algorithmic regulation and the rule of law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018;376(2128):20170355.

[19] Cornelissen NAJ, van Eerdt RJM, Schraffenberger HK, Haselager WFG. Reflection machines: increasing meaningful human control over Decision Support Systems. Ethics and Information Technology. 2022;19(24).

[20] Cabitza F, Campagner A, Sconfienza LM. Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading. Health Information Science and Systems. 2021;9(1):1-20.

[21] Kasparov G, Greengard M. Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins. Millennium Series. John Murray Press; 2017. Available from: `https://books.google.it/books?id=ffYZDQAAQBAJ`.

[22] 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. 2021;155:102696.

[23] Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, et al. The ethics of algorithms: key problems and solutions. AI & SOCIETY. 2021:1-16.

[24] Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. Journal of medical ethics. 2020;46(3):205-11.

[25] Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. Jama. 2017;318(6):517-8.

[26] Bond RR, Novotny T, Andrsova I, Koc L, Sisakova M, Finlay D, et al. Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. Journal of electrocardiology. 2018;51(6):S6-S11.

[27] Spitzer M. Outsourcing the mental? From knowledge-on-demand to Morbus Google. Trends in Neuroscience and Education. 2016;5(1):34-9.

[28] Carr NG. The Shallows: What the Internet is Doing to Our Brains. W.W. Norton; 2010. Available from: `https://books.google.it/books?id=9-8jnjgYrgYC`.

[29] Malone TW. Superminds: The Surprising Power of People and Computers Thinking Together. Little, Brown; 2018. Available from: `https://books.google.it/books?id=Qe0zDwAAQBAJ`.

[30] Pierce J. Undesigning interaction. Interactions. 2014;21(4):36-9.

[31] Cabitza F. Biases affecting human decision making in AI-supported second opinion settings. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer; 2019. p. 283-94.

[32] Boni M. The ethical dimension of human–artificial intelligence collaboration. European View. 2021;20(2):182-90.

[33] Taddeo M, Floridi L. How AI can be a force for good. Science. 2018;361(6404):751-2.

[34] Floridi L. Establishing the rules for building trustworthy AI. Nature Machine Intelligence. 2019;1(6):261-2.

[35] Methnani L, Aler Tubella A, Dignum V, Theodorou A. Let Me Take Over: Variable Autonomy for Meaningful Human Control. Frontiers in Artificial Intelligence. 2021;4. Available from: `https://www.frontiersin.org/article/10.3389/frai.2021.737072`.

[36] of Europe's Ad Hoc Committee on AI (CAHAI) TC. Towards Regulation of AI Systems. Council of Europe; 2020.

[37] on Artificial Intelligence HLEG. Ethics Guidelines for Trustworthy Artificial Intelligence. European Commission; 2019.

[38] Shahriari K, Shahriari M. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC). IEEE; 2017. p. 197-201.

[39] European Commission, Joint Research Centre, Nativi S, De Nigris S. AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework. Publications Office; 2021.

[40] Smuha NA. The EU approach to ethics guidelines for trustworthy artificial intelligence. Computer Law Review International. 2019;20(4):97-106.

[41] Meza Martínez MA, Nadj M, Maedche A. Towards an integrative theoretical framework of interactive machine learning systems. 2019.

[42] Wiethof C, Bittner EA. Hybrid Intelligence–Combining the Human in the Loop with the Computer in the Loop: A Systematic Literature Review. 2021.

[43] Xu W, Dainoff MJ, Ge L, Gao Z. From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI. ArXiv. 2021;abs/2105.05424.

[44] Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics. 2016;3(2):119-31.

[45] Parasuraman R, Manzey DH. Complacency and bias in human use of automation: An attentional integration. Human factors. 2010;52(3):381-410.

[46] Shneiderman B. Human-centered artificial intelligence: three fresh ideas. AIS Transactions on Human-Computer Interaction. 2020;12(3):109-24.

[47] Skitka LJ, Mosier K, Burdick MD. Accountability and automation bias. International Journal of Human-Computer Studies. 2000;52(4):701-17.

[48] Sloman S, Fernbach P. The Knowledge Illusion: Why We Never Think Alone. Penguin; 2017.

[49] Frischmann B, Selinger E. Re-engineering humanity. Cambridge University Press; 2018.

[50] Scalera L, Gallina P, Gasparetto A, Seriani S. Anti-Hedonistic Machines. Int J Mech Control. 2017;18:9-16.

[51] Cabitza F, Campagner A, Ciucci D, Seveso A. Programmed inefficiencies in DSS-supported human decision making. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer; 2019. p. 201-12.

[52] Ohm P, Frankle J. Desirable inefficiency. Fla L Rev. 2018;70:777.

[53] Tenner E. The Efficiency Paradox: What Big Data Can't Do. Knopf Doubleday Publishing Group; 2018. Available from: `https://books.google.it/books?id=PgAtDwAAQBAJ`.

[54] Hildebrandt M. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. Theoretical Inquiries in Law. 2019;20(1):83-121.

[55] Sadin É. L'intelligence artificielle, ou, L'enjeu du siècle: anatomie d'un antihumanisme radical. Collection Pour en finir avec. L'Échappée; 2018. Available from: `https://books.google.it/books?id=yJ1uvQEACAAJ`.

[56] Muller JZ. The Tyranny of Metrics. Princeton University Press; 2018. Available from: `https://books.google.it/books?id=J3GYDwAAQBAJ`.

[57] Crawford K. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press; 2021. Available from: `https://books.google.it/books?id=KfodEAAAQBAJ`.

[58] Buçinca Z, Malaya MB, Gajos KZ. 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. 2021;5(CSCW1):1-21.

[59] Christakis NA. Blueprint: The Evolutionary Origins of a Good Society. Little, Brown Spark; 2019.

[60] Park JS, 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. 2019;3(CSCW):1-15.

[61] Holzinger AT, Muller H. Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI. Computer. 2021;54(10):78-86.

[62] Kaur S, Sharma R. Emotion AI: Integrating Emotional Intelligence with Artificial Intelligence in the Digital Workplace. In: Singh PK, Polkowski Z, Tanwar S, Pandey SK, Matei G, Pirvu D, editors. Innovations in Information and Communication Technologies (IICT-2020). Cham: Springer International Publishing; 2021. p. 337-43.

[63] Jankuloski F, Bozinovski A, Pacovski V. Artificial Intelligence: Simulating Human Emotion and Surpassing Human Intelligence. 2020.

[64] Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. Ai & Society. 2021:1-7.

[65] Nass C, Steuer J, Tauber E, Reeder H. Anthropomorphism, agency, and ethopoeia: computers as social actors. In: INTERACT'93 and CHI'93 conference companion on Human factors in computing systems; 1993. p. 111-2.

[66] Hamacher A, Bianchi-Berthouze N, Pipe AG, Eder K. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE; 2016. p. 493-500.

[67] Shneiderman B. Human-centered AI. Issues in Science and Technology. 2021;37(2):56-61.

[68] Hildebrandt M. New animism in policing: re-animating the rule of law. The SAGE handbook of global policing. 2016:406-28.

[69] Atkinson C, Brooks L. In the Age of the Humanchine. ICIS 2005 Proceedings. 2005:11.

[70] IEEE. Ethically Aligned Design (Version 1). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2016.

[71] Turing AM. Computing Machinery and Intelligence. Mind. 1950;59(October):433-60.

[72] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3(11):e745-50.

[73] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence

(XAI). IEEE Access. 2018;6:52138-60.

[74] Cabitza F, Alderighi C, Rasoini R, Gensini GF. "Handle with care": about the potential unintended consequences of oracular artificial intelligence systems in medicine. Recenti progressi in medicina. 2017;108(10):397-401.