# AI Shall Have No Dominion: on How to Measure Technology Dominance in AI-supported Human decision-making


Federico Cabitza*
University of Milano-Bicocca
Milan, Italy
IRCCS Ospedale Galeazzi - Sant'Ambrogio
Milan, Italy
federico.cabitza@unimib.it

Andrea Campagner*
IRCCS Ospedale Galeazzi - Sant'Ambrogio
Milan, Italy
andrea.campagner@unimib.it

Riccardo Angius
University of Padova
Padua, Italy

Chiara Natali
University of Milano-Bicocca
Milan, Italy

Carlo Reverberi
University of Milano-Bicocca
Milan, Italy



## ABSTRACT

In this article, we propose a conceptual and methodological framework for measuring the impact of the introduction of AI systems in decision settings, based on the concept of *technological dominance*, i.e. the influence that an AI system can exert on human judgment and decisions. We distinguish between a negative component of dominance (automation bias) and a positive one (algorithm appreciation) by focusing on and systematizing the patterns of interaction between human judgment and AI support, or *reliance patterns*, and their associated cognitive effects. We then define statistical approaches for measuring these dimensions of dominance, as well as corresponding qualitative visualizations. By reporting about four medical case studies, we illustrate how the proposed methods can be used to inform assessments of dominance and of related cognitive biases in real-world settings. Our study lays the groundwork for future investigations into the effects of introducing AI support into naturalistic and collaborative decision-making.



**ACM Reference Format:**
Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. 2023. AI Shall Have No Dominion: on How to Measure Technology Dominance in AI-supported Human decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3544548.3581095


## 1 INTRODUCTION

The use of artificial intelligence (AI) systems for decision support and task automation has recently gained great popularity both in scientific and institutional contexts and in public opinion, reaching the level where it is considered natural, and almost a need, to adopt these systems even in areas where decisions have a legally relevant impact on those involved and the error rate or arbitrariness of decision-makers is considered unacceptable [55] (such as in medicine, court decisions, public safety, credit-worthiness). This interest and normalization are largely based on the frequently unstated presumption that the fewer mistakes an AI support makes, the better it is [13]. The appeal of this assumption is largely connected to its simplifying consequences: to decide whether an AI system is good enough for deployment in practical contexts it suffices to evaluate its performance in isolation, or perhaps by comparing its performance with the average performance of human decision-makers in the same task [4, 69], without necessarily paying consideration to the complex socio-technical context [32] in which the system itself will be embedded after deployment. The point is that this dogma is appealing in theory as it is of low applicability "in the wild" [56], as it is reasonable only in regard to the small number of cases in which humans adopt a fully automated decision-making setting and completely delegate decision-making to machines [3, 55]. However, in the overwhelming majority of cases, the automation of classifying tasks is partial [75], and intended as a support to the human decision, i.e. to an act factually performed by a human being and for which they are solely responsible [1]. In these contexts, that can be framed under the definition of *human-centered AI* [83], the above-mentioned assumption is not only inapplicable but also harmful [16]; in all those cases, indeed, the evaluation of an AI system should be aimed at understanding its role in letting people either avoid or commit incorrect decisions, by factoring in both cognitive and socio-psychological determinants and effects [51, 65].

But let us briefly move to a more general level of discourse, where we abstract from any specificity of AI systems: automation is the allocation of functions to technology, that is justified in all those situations where it may improve performance, usually in terms of efficiency (e.g., faster processing) and safety (e.g., fewer

*Both authors contributed equally to this research.



*CHI '23, April 23–28, 2023, Hamburg, Germany*



ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581095


[1] We note that in these cases scholars speak, with an abuse of language, of human-in-the-loop [49] rather than adopt other terms such as computer-in-the-group [83]. See also Figure 3.

incidents), possibly also in terms of effectiveness (e.g., fewer errors), and also in regard to user satisfaction or comfort (especially for non-mandatory systems that users have to purchase to use). More specifically in decision support, technology is usually applied under the assumption jokingly known as the *fundamental theorem of Informatics* [36]: the use of technology must improve the unaided human (i.e., $H + A > H$). Also taking into account the literature concerned with the use of AI systems as tools for decision support, this idea is not new, although seldom mentioned. Indeed, borrowing from the health informatics domain, in particular from the CONSORT AI [60] reporting guideline, we recall two relevant concepts that have been recently proposed to formalize the above-mentioned intuitive perspective: first, the notion of an *AI intervention*, i.e. any intervention which relies on an artificial intelligence/machine learning component to serve its purpose; second, the notion of a *human-AI interaction protocol* [15], i.e. the process stipulating how humans interact with the affordances mobilized in the AI intervention, for this latter to function as intended and achieve its objectives.

In analogy with the use of the term in the health sciences literature, the above definitions allow us to frame decision support through AI automation as an intervention: thus, an AI intervention must have a practically relevant (not just statistically significant) effect or impact on human decision-making to make sense, that is to be worth the effort as well as the risk of occurring in some (substantially unavoidable) side effect or unintended consequence [17, 67].

There is a wide and inter-disciplinary literature that has focused on the side effects of computationally-supported decision-making, which are often referred to with expressions such as loss of situational awareness, automation complacency and bias, and skill degradation [17, 39]. Within this wide and varied literature, we chose to focus on the *theory of technology dominance* [5]: this is a framework that was proposed to consider the main determinants that make users more or less reliant on decision aids [53], as well as the main facilitating factors for the occurrence of the unintended effects mentioned above. Within this theoretical framework, two main constructs are defined: *reliance*, that is, quite obviously as the term suggests, "the extent to which an individual uses the intelligent decision aid and integrates the recommendations of that aid into their judgment" [5]; and *dominance*, which is a much more interesting and less investigated concept. The authors defined dominance as the "the state of decision-making where the intelligent decision aid, as opposed to its user, takes primary control of a decision-making process" [5]. The former concept of reliance has so far garnered far more attention as a result of this radical position [87], and the scope of domination may have appeared to be limited to the higher (and less probable) levels of automation [75] where the "computer decides everything, acts autonomously, [even possibly] ignoring the human or informing the human only if, the computer, decides to". Nonetheless, other authors [86, 87] have recently related the concept of dominance to a more common situation: "the dominating influence that technology may have over the user, which allows the user to take a more subservient position—in essence, the user deferring to the technology in the decision-making process." [87].

We will take this latter definition for our idea of AI dominance in decision-making. Indeed, we believe that studies about the effectiveness of the allocation of decision functions to machines should focus on their *influence* on the user's judgment and discretion [86], with interesting and partly unexplored links with the research conducted under the *CAPTology* label [34, 35]. The main motivation for our approach, then, lies in the following consideration: while decision support can be found useful also when it limits itself in confirming its user's decision, in the role of double-checker and confidence-builder, the value of its output (advice) lies in the *incremental probability* that the *best* option, i.e., the one with the most positive consequences (e.g., the improvement of one's health conditions) will be chosen and finally enacted; or put even more strongly, in the fact that a chain is started where the initial human decision is changed, the subsequent process altered and the related outcome improved [22]. In other words, *automation utility lies in the influence and dominance* that it exercises when the advice and recommendations that it produces are correct: that is, a successful AI intervention is one where the system influences and persuades users (for the better), what in literature has also been called (beneficial) *algorithm appreciation* [50]. As a matter of fact, so far we have focused on positive technology dominance, when the machine helps users avoid a mistake they would have made without receiving its advice; however, and obviously so, dominance can also be negative, when the machine misleads the user and causes them to make a mistake or overlook a situation. In the specialist literature this is the case for automation complacency (for omissions) and automation bias (commission errors). Strong evidence suggests that one cannot have positive dominance without some portion of the negative one, although research has been aimed at minimizing the occurrence and impact of the latter one. In fact, we recognize that the theory of technology dominance mentioned above is one of the research initiatives more clearly and programmatically aimed at mitigating the risk of over-reliance in decision support users [5, 88].

In this article, we will focus on the *fit* between human decision makers and AI support in classification tasks, that is on the 'F' ambit of the framework we denoted as *AFOOT* and briefly outlined in [15]. In particular, the main contribution of this article is the proposal of a general methodological framework, which encompasses both simple metrics as well as practical visualizations, to gauge *technology dominance*. To this aim, we will distinguish between positive and negative dominance, and related biases, such as automation bias, algorithm appreciation, and a phenomenon that deserves more attention: the *white box paradox* (shortly put, whenever automation bias is influenced by the provision of explanations). We believe that this latter phenomenon is particularly relevant in light of the recent interest towards eXplainable AI (XAI), that is methods and techniques aimed at making the output of an AI intervention more understandable to the users. We will then illustrate the use of the proposed methods by applying them in four user studies that we conducted in the last months involving tens of physicians in simulated diagnostic tasks.

## 2 RELATED WORK

The following is a brief but hopefully comprehensive review of the positions expressed in the literature on the two main and opposing manifestations of technological dominance in decision-making contexts: what are denoted in some references as under- and

over-reliance (e.g. [80, 82][2]), and in others as automation aversion and automation appreciation (e.g., [27, 61]), and on the risks associated with these opposing attitudes, including, above all, automation bias [64]. A summary of the concepts and metrics discussed in this section is shown in Table 1.

## 2.1 On the theory of technology dominance

As briefly recalled in the introduction, the theory of technological dominance has originally been proposed by Arnold and Sutton [5] as a framework to consider the main determinants that influence users' reliance on decision aids, as well as to characterize the factors leading to over-reliance, dependence and deference. Since its proposal, the theory of technological dominance has been explored in several user studies, starting from the first decade of the 2000s. Most work in this sense has been conducted in the organizational domain, for example in the the fields of accounting [72], tax reporting [66], credibility assessment [53] and business intelligence [91]. These works are characterized by the use of qualitative methods (observations, interviews, questionnaires), mostly aimed at identifying and characterizing the determinants (i.e., facilitating factors) of technology dominance: for instance, Noga and Arnold [72] identified the level of expertise of the users as one of the main determinants of dominance, showing that less expert users were more easily prone to dominance than the more experienced ones; on the other hand, Hampton and Williams [45, 91] identified the complexity of the technological aids as a distinct, and also relevant, determinant of dominance. More recently, Sutton et al. [88] proposed an extensions of the original theory in which they explicitly relate it to two psychological phenomena that have recently attracted the interest of the human-AI interaction research community, namely *algorithm appreciation* and *aversion*, which therein are characterized as being two opposite dimensions on the *reliance/non-reliance* spectrum of dominance. As they stand, both algorithm appreciation and aversion conceptualise advice utilization (or lack thereof), i.e. how much, relative to the specific unit of measurement required by a task, the original decision is changed to match algorithmic advice, and do not concern the quality of final decisions. The authors also explicitly discuss one of the most studied cognitive effects emerging in the decision support domain, namely *automation bias* which they identify as a possible detrimental effect of algorithm appreciation, when this latter causes people to be less vigilant and exert less situational awareness [85].In the following we will build on the recent framework of Sutton et al. to flesh out our proposal, in particular we will adopt the notion of algorithmic aversion as corresponding to non-reliance on AI intervention, while, for clarity, we will use the terms automation bias and algorithm appreciation to denote, respectively, the negative and positive dimensions of reliance. Nonetheless, before going to the core of our proposal, we will briefly review the literature centered around these three notions.

[2]In their framework, Schemmer et al [82] distinguish different forms of reliance: *Positive AI Reliance*: wrong initial decision, correct AI advice, correct final decision. *Negative Self-reliance*: wrong initial decision, correct AI advice, wrong final decision. *Positive Self-reliance*: correct initial decision, wrong AI advice, correct final decision. *Negative AI-reliance*: correct initial decision, wrong AI advice, wrong final decision.

## 2.2 Algorithm appreciation, aversion and related metrics

In regard to the notion of algorithm appreciation, this has been defined by Logg et al. [61] as the condition where "people consistently give more weight to equivalent advice when it is labeled as coming from an algorithmic versus human source", which in the theory of technological dominance can be equated to automation (over-)reliance [88]. As said above, studies within the technology dominance mold mostly focus on the identification and assessment of the determinants of automation appreciation,while the definition of quantitative metrics to evaluate this phenomenon has hardly been considered in the related literature. The most relevant contribution in this sense has been proposed by Logg et al. [61], who proposed the use of a metric, called *Weight of Advice* (WOA) and originally proposed in the organizational setting [46, 47], to measure the extent to which algorithm appreciation occurs in a decision setting [9]. This latter one is defined as "the difference between the initial and revised judgment divided by the difference between the initial judgment and the advice", meaning "WOA of 0% occurs when a participant ignores advice and WOA of 100% occurs when a participant abandons his or her prior judgment to match the advice.". Notably, since it is defined in terms of differences and ratios among raw judgments' values, the WOA can be applied only to continuous and binary judgments, noticing that in the binary case the metric can be undefined whenever the prior judgment and the advice coincide. Furthermore, the WOA does not provide any indication in regard to the correctness of the expressed judgments, but only a measure of agreement between the human judgment and the advice.

Intuitively related to algorithm appreciation, the notion of algorithmic aversion has been defined by Dietvorst et al. in [27]: the authors, by drawing inspiration from previous studies in the forecasting setting, investigated conditions which may explain why "when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster" although "research shows that evidence-based algorithms more accurately predict the future than do human forecasters". Indeed, as mentioned before and also discussed in [54, 88], the notion of algorithmic aversion can be understood as the opposite of algorithm appreciation, and put into correspondence with non-reliance on the decision support or absence of dominance exercised by this latter.In regard to the measurement of aversion, a metric to this purpose, called Shift and originally proposed in the field of forecasting to measure advice utilization [73], is proposed by Prahl and Van Swol in [79]. Shift is a variant of the WOA discussed above, and can similarly be computed as the difference between the revised and initial judgment divided by the difference between the advice and the initial judgment. Obviously, since this is equivalent to the WOA, the same considerations and limitations described above for this latter metric also apply to Shift.

## 2.3 Automation bias and related metrics

Finally, in regard to automation bias, this is, among the shortcomings related to automation (which include phenomena such as situational awareness, complacency, skill degradation) and associated with the notion of dominance, one of the more frequently observed

**Table 1: Summary of the metrics discussed in related work**

| Metric | Abbrev. | Definition | Source |
|---|---|---|---|
| AI Effect on Final Decision | $\omega_I$ | $\frac{\text{odds(the final decisions were the same as AI advice)}}{\text{odds(the initial decisions were the same as AI advice)}}$ | Reverberi et al., 2022 [80] |
| AI Effect on Accuracy | $\omega_A$ | $\frac{\text{odds(the final decisions were correct)}}{\text{odds(the initial decisions were correct)}}$ | |
| AI Effectiveness | $\omega_E$ | same as $\omega_A$, but using odds estimated only on the subset of cases where AI advice was correct | |
| AI Safety | $\omega_S$ | same as $\omega_A$, but using odds estimated only on the subset of cases where AI advice was wrong | |
| Shift (advice utilisation) | n.a. | $\frlvert\text{Final decision - Initial Decision}\frrvert$ / $\lvert\text{Advice - Initial Decision}\rvert$ | Prahl and Van Swol, 2017 [79] |
| Weight of Advice (advice utilisation) | WOA | $\frlvert\text{Final Decision - Initial Decision}\frrvert$ / $\lvert\text{Advice - Initial Decision}\rvert$ | Logg et al., 2019 [61] |
| Automation Bias | *AB* | % of cases where accuracy worsened after AI advice | Goddard et al., 2014 [39] |
| Net Improvement | NI | $\left(\begin{array}{l}\textit{\% of cases where}\\ \textit{accuracy improved}\\ \textit{after AI advice}\end{array}\right) - \left(\begin{array}{l}\textit{\% of cases where}\\ \textit{accuracy worsened}\\ \textit{after AI advice}\end{array}\right)$ | |
| Relative Positive AI Reliance | *RAIR* | $\frac{\text{Positive AI Reliance}}{\text{Positive AI Reliance + Negative Self-Reliance}}$ | Schemmer et al., 2022 [82] |
| Relative Positive Self-Reliance | *RSR* | $\frac{\text{Positive Self-Reliance}}{\text{Negative AI Reliance + Positive Self-Reliance}}$ | |

and focused upon ones. Automation bias has so far been defined in many scholarly contributions, and even in an International Standard: the "type of human cognitive bias due to over-reliance on the recommendations of an AI system"[3]. In the last 10 years, two systematic reviews on the topic have been published, although the latest one was written more than 5 years ago: in these, and other, contributions, automation bias has been defined in various ways, which nevertheless share the same basic idea mentioned above. In particular, automation bias has been defined as "automation-included complacency" in [63], or, more specifically, as "the propensity of people to over rely on automated advice" [39], and as the "tendency to over-trust HIT [Healthcare Information Technology] leading a physician to make an incorrect decision in order to follow the advice provided by a CDSS [Clinical Decision Support System]" in [11]; also as "the human tendency … which occurs when a human decision maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct" by [23]; and, finally, as the tendency "by which users tend to over-accept computer output 'as a heuristic replacement of vigilant information seeking and processing"'. This latter way to define automation bias is an oft-cited definition by Mosier and Skitka from [71], which has also been referenced in connection with the theory of technology dominance [88] and which we will also take as our main reference in the following. Although the existence of automation bias has been widely established, and various research works have developed qualitative approaches aimed at identifying determinants of automation bias [2, 31, 70], efforts to provide quantitative and analytical evaluation of the impact of this bias on decision-making are still lacking. Indeed, the most recent systematic review, by Lyell and Coiera ([63]) in 2016, illustrated that "Only 9 studies reported the significance for the presence of automation bias compared to a manual (nonautomated) control. This, combined with the large variability in the reported measures, makes it difficult to draw comparisons between studies.". Our own investigation on the usage of metrics in ensuing studies, which we briefly review in the following, and illustrated in Table 1, appears to confirm this is still the case and the field is yet to settle on the definition of rigorous metrics. One of the most significant contributions in this sense is the work by Goddard et al. in [39], who measure automation bias as the rate of "decision switches from correct pre-advice, to incorrect post-advice", analysing its occurrence in the context of

[3] cf. ISO/IEC TR 24027:2021, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision-making.

a decision support system (DSS) for medication prescription.They also put forward an approach to measure the *net improvement* to assess the DSS performance, taken as the difference between the percentage of cases where prescription accuracy improved and automation bias, i.e. the percentage of cases where prescription accuracy worsened. The same approach proposed by Goddard et al. has also been considered in other domains, e.g. in medicine [11], in airport security [24], or in child welfare [25].

## 2.4 Further related work on the measurement of dominance

Concluding this section, we recall two frameworks, proposed by Schemmer et al. in [82] and Reverberi et al. [80], which while not directly proposed within the context of the theory of technology dominance, share some features with, and provide a grounding step to, our proposal. Schemmer et al. proposed a conceptual framework to measure *positive reliance* i.e. "the human's ability to differentiate between correct and incorrect AI advice and to act upon that discrimination". This is achieved by first proposing a characterization of four different dimensions of reliance (what in the following we call *reliance patterns*) which are then used to define two metrics, *relative positive-self reliance* and *relative positive reliance*, which indicate the level of beneficial reliance of the human decision-maker respectively on themselves or on the AI and are thus directly related to the positive dimension of reliance and non-reliance. The authors also propose qualitative visualizations to graphically evaluate the mentioned metrics, and discuss the use of the proposed framework to determine whether changes to an AI system (e.g. the adoption of Explainable AI methods) may result in incremental improvements towards appropriate reliance and better human-AI teaming performance in turn. While the terminology used by the authors does not directly mention neither dominance nor its positive and negative aspects, it is not hard to see that the notion of reliance can be directly matched with that of dominance, while *relative positive-self reliance* and *relative positive reliance* can be understood as metrics for quantifying the extent to which an AI intervention induces on the decision-makers, respectively, a beneficial algorithmic aversion (i.e. the absence of negative dominance) and algorithm appreciation (i.e. positive dominance). Reverberi et al., by comparing odds of outcomes with and without the support of an AI intervention, propose several odds ratio-based metrics, of which two (i.e. *effectiveness* and *safety*) are directly related, respectively, to positive reliance (i.e. a beneficial form of algorithm appreciation) and absence of negative reliance (which, in turn, can be related to automation bias). The same metrics are applied by the authors in evaluation of the usefulness of introducing an AI intervention in a medical decision-making setting.

# 3 DIMENSIONS OF THE EFFECTS OF AI INTERVENTIONS

If, as mentioned in the introduction, AI is not seen as an agent [83], but rather as an intervention, a natural way to assess its impact on work practices and decision tasks is on a comparative basis. This is routinely done, for instance, in various high stake domains, such as healthcare, where medical interventions, such as surgical operations or drugs, are compared by adopting in each case the method best suited to the circumstances and available resources (from cohort observational studies to more demanding RCTs). Comparisons can be longitudinal, that is along the temporal dimension: this entails comparing the performance of a human settings (at any granularity level, from the small team to the business unit or whole organization) *before* and *after* technology adoption over time [19], assuming that no other relevant change (with respect to performance) has occurred in the same lapse of time. Comparisons can also be cross-sectional: this entails comparing the performance of two settings, one *aided* and the other one *unaided*, which, excluding the available technology under test, should be equivalent for any practical comparative purposes.

In comparative terms, many ways to measure accuracy improvement (or degradation) are possible. The odd ratio of accuracy scores (or its complement, error rate) is probably the more straightforward: this measure considers the probability (i.e., odd) to make a decision error with AI with respect to the probability to err without the AI. By definition, odd ratios greater than one are indicative of settings where the AI intervention improves accuracy with respect to the control, unaided, group. Particular care must be paid to ensure that the conditions are truly comparable, e.g. the complexity and nature of the cases considered either, as well as the skills of the people involved (as it is regularly done in prospective, randomised and controlled experimental settings).

In what follows, we will focus on the impact that an AI intervention has on decision autonomy, what in Section 1 we called dominance, that is the influence that makes humans change their minds (as well as their decisions): this latter definition of dominance makes clear that the assessment of dominance requires a specific focus on longitudinal case studies (specifically so, cohort studies), rather than cross-sectional ones, so as to allow potential changes of decision under the influence of an AI intervention. To illustrate this dimension and be able to detect and characterize its components, it is useful to introduce the idea of a *decision table* (See 2), by which all possible combinations of the unaided (or initial) human judgment, the AI support and the final decision made by the user in light of the decision aid are enumerated: each of these three judgments can be either correct (1) or wrong (0) with respect to a reference ground truth (which is assumed to be totally correct and reliable). The notion of decision table directly draws from the partial enumeration of the effects of AI advice on human reliance proposed by Schemmer et al. [82]. With respect to this first contribution, we enumerate four additional reliance patterns (namely, detrimental reliance, beneficial under-reliance, detrimental self-reliance and beneficial over-reliance) which were not considered in [82] where the authors focused mostly on the positive side of reliance and non-reliance. Furthermore, we also associate to the reliance patterns the cognitive biases and effects as related to theory of technology dominance.

As can be easily observed from the table [82], and as already discussed in the previous sections, an AI intervention can have two main effects on decision accuracy, a positive and a negative one. In its turn, the positive effect has two components:

(1) P1. When the AI helps users to confirm their initial right judgment (see pattern 111 in Table 2.);

**Table 2: Definition of all possible decision and reliance patterns between human decision makers and their AI system. In the first three columns, 0 denotes an incorrect decision, and 1 a correct decision. We associate the attitude towards the AI in each possible decision pattern (in terms of trust [52]), which leads to either accepting or discarding the AI'advice, and the main related cognitive biases.**

| **Human judgment** (H) | **AI support**[4] (AI) | **Final decision** (D) | **Reliance pattern** | **Biases and Effects** |
|---|---|---|---|---|
| 0 | 0 | 0 | detrimental reliance | automation complacency |
| 0 | 0 | 1 | beneficial under-reliance | extreme algorithmic aversion |
| 0 | 1 | 0 | detrimental self-reliance | conservatism bias |
| 0 | 1 | 1 | beneficial over-reliance | algorithm appreciation |
| 1 | 0 | 0 | detrimental over-reliance | automation bias |
| 1 | 0 | 1 | beneficial self-reliance | algorithmic aversion |
| 1 | 1 | 0 | detrimental under-reliance | extreme algorithmic aversion |
| 1 | 1 | 1 | beneficial reliance | confirmation bias (in later cases) |

(2) P2. When the AI helps users to avoid their initial wrong judgment (011). This case is also known as (beneficial) *algorithm appreciation*: as mentioned previously, for simplicity, we will denote this case as algorithm appreciation.

For an AI intervention to be useful and beneficial (and not detrimental) the positive effect must be greater than the negative effect, which, similarly to the positive one, has two components:

(1) N1. When the AI fails to make users change their minds about wrong judgments (010);
(2) N2. When the AI AI misleads users by inducing otherwise avoidable errors (100). This notorious case is also known in the specialist literature as *automation bias*.

Clearly, P2 and N2 are the main factors in AI dominance, as they relate to the power of changing the mind of the decision-maker, altogether with much rarer cases of human rejection of the AI advice that lead to changes of judgements and that we associated with extreme algorithmic aversion (i.e. patterns 001 and 110 in Table 2). Based on the relative frequency with which these patterns can be observed in a practical settings, different forms of dominance can be distinguished: in what follows we introduce metrics to quantify these different forms of dominance, focusing on, respectively, algorithm appreciation, automation bias and (detrimental) algorithmic aversion.

# 4 METHODS AND METRICS FOR ASSESSING DOMINANCE

As mentioned above, in the following we will describe two different approaches to quantify the effect of AI interventions in a decision setting, and the dominance it exerts. First, a *frequentist* approach, which solely considers the ratios and frequencies appearing in a study's decision table; then, a more general *causal* approach, which augments the frequentist one by considering the causal history of the decision-making process, thus allowing to obtain a more reliable and generalizable estimate of the causal effect of the AI intervention.

## 4.1 The Frequentist Approach

As mentioned above, the first approach, which we call *frequentist* since it is based solely on the frequencies drawn from the decision table, starts from the measuring of the number of errors made with and without the support of AI. These quantities, which are derived from standard epidemiological definitions, are computed to quantify the effect (if any) of the AI intervention, and whether this was on average positive or negative. We consider, in particular, the following rates:

$$AIE = \frac{\text{number of cases where } D = 0}{N},$$
$$CE = \frac{\text{number of cases where } H = 0}{N},$$
$$AIN = \frac{\text{number of cases where } D = 1}{N},$$
$$CN = \frac{\text{number of cases where } H = 1}{N},$$

where $N$ is the total number of cases. Intuitively, the above quantities represent the (relative) number of errors made by humans when aided by an AI intervention (AIE); the (relative) number errors made when unaided, that is without the AI support (CE, this is the number of errors made in the Control group in cross-sectional settings); the (relative) number of right AI-aided decisions (AIN); and the (relative) number of right unaided decisions (CN). In Appendix A we describe procedures to compute the above quantities in terms of the reliance patterns mentioned in Table 2.

When the 4 previously described amounts are known, two conditional error rates can be defined, namely:

- The Error Rate conditioned on being aided: AIER = AIE / (AIE+AIN);
- The Error Rate conditioned on being unaided: CER = CE / (CE + CN).

From these two error rates, the utility of the AI intervention can be quantified by means of the odds ratio:

$$\text{Technology Impact } = \frac{CER}{1 - CER}\frac{1 - AIER}{AIER}$$

The Technology Impact (TI) is the reciprocal of the "effect on diagnostic accuracy" defined in [80]. Compared to this latter metric, the interpretation of the TI is more similar to its counterpart in epidemiological studies: the TI is the representation of the ratio of the likelihood of an error in the supported group with respect to the probability of an error in the unsupported one. Obviously, values larger than 1 indicate an overall positive effective (that is the AI intervention had a pragmatically useful effect), suggestive of the fact that the positive component of dominance outweighs the negative one; conversely, values below 1 denote a detrimental effect of the AI intervention on decision-making.

Having quantified the general effect of the AI intervention, the frequentist approach can be used also to assess the positive and negative components of dominance induced by the intervention, focusing thus on automation bias (i.e. the negative component of dominance) and algorithmic appreciation (i.e. the positive component of dominance), or its complement, (detrimental) algorithmic aversion. Then, we define two odds ratio-based metrics that quantify the automation bias and the detrimental algorithmic aversion:

$$\text{Automation Bias} = \frac{dor}{N - dor}\frac{N - bsr}{bsr}$$

$$\text{Detrimental Algorithmic Aversion} = \frac{dsr}{N - \text{dsr}}\frac{N - bor}{bor}$$

where *dor* is the detrimental over-reliance, *bsr* the beneficial self-reliance, *dsr* thedetrimental self-reliance and *bor* the beneficial over-reliance. Intuitively, the *automation bias odds ratio* is defined as the ratio between the likelihood of automation bias (associated with detrimental over-reliance) and that of beneficial algorithmic aversion (associated with beneficial self-reliance): thus, values larger than 1 hint at the fact that the AI intervention may induce automation bias and thus exert negative dominance, misleading the users in following an incorrect advice (negative effect N2); on the other hand, values below 1 suggest the absence of such a negative effect. Similarly, the *detrimental algorithmic aversion odds ratio* is defined as the ratio between the likelihood of conservatism bias (associated with detrimental self-reliance) and that of algorithmic appreciation (associated with beneficial over-reliance): values larger than 1 hint at the fact that the AI intervention fails at exerting positive dominance (positive effect P2), with users failing to change their minds about their wrong decisions, while values below 1 hint at the opposite positive effect determined by algorithmic appreciation.

## 4.2 The Causal Approach

The frequentist approach outlined above quantifies the effect of an AI intervention based on the *associations* between human and AI decisions in a given decision-making setting, as observed in the corresponding decision table: the term *association* here refers to the fact that the frequentist approach ignores the *causal history* of the decision problem, i.e. how the human judgments' and AI suggestions were produced and how they interacted to inform the final human decision, since it only considers the observed correlations among frequency rates (re Pearl's *ladder of causation* [78]). In what follows, we consider an alternative and more general approach to quantify the *causal* dominance effect exerted by an AI intervention in decision settings, based on causal inference methods within the Halpern-Pearl framework [44]. To illustrate this approach, let

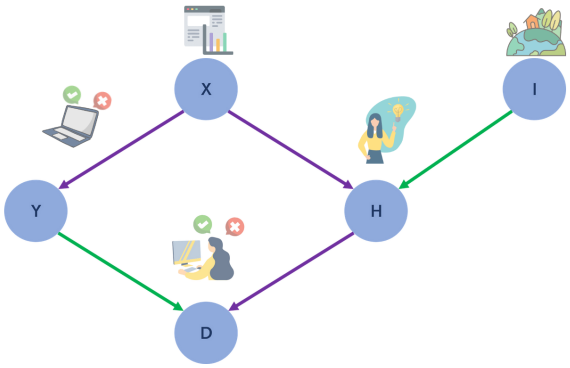


**Figure 1: Causal diagram for Human-AI interaction, generated with DAGitty. Each node represents a random variable, which can be either observed or latent (non-observed). Arcs denote dependence relationships among random variables: in particular, groups of incoming arcs define a functional relationship (i.e. a random variable value is defined as a non-deterministic function of its parents' nodes). The D node is the outcome, that is the final decision; Y is the exposure, that is the machine's advice (i.e., output of its functional model); X denotes all information available in digital format and input to the ML model (which possibly includes personalization information or some user characteristics); I is a latent *context* variable; H is the initial, unaided classification by the human. Transitive arcs (i.e., the arcs from X to D and from I to D) are not explicitly depicted. Green arrows denote the causal path, while purple arrows denote a biasing (backdoor) path. The causal path denotes the sequence of relationships (in this case, the arc connecting Y to D) which represents the causal effect of interest. A biasing, or backdoor, path represents an alternative path which connects the cause and the effect variables of interest by passing through a confounder or common cause (in the Figure, H).**

us assume that the *causal history* which generated an observed decision table can be represented by the causal diagram depicted in Figure 1, which describes the potential interaction between a human decision-maker and an AI intervention.

Thus, we assume that the human decision maker forms an initial judgment (H), unaided by the AI intervention, by considering all the available information about a case, that is both what is available in digital format (X) and what is not (I), i.e. what is not observable by the machine. Then the human decision maker receives the machine's advice (Y), which in turn is based on X, and finally reaches and produces a final decision (D), where we assume that this final decision is arrived at by only considering the machine' advice and their initial judgment (that is without considering any new evidence about the case), as well as X and I. More in detail, Y represents the advice of the AI intervention (more precisely, whether the advice is correct or wrong) while H represents the first opinion formulated by the human decision maker (or, more precisely, whether this latter opinion would be correct or wrong). Both H and Y are influenced by X, that is the characteristics (or features) of the case at hand that are available to both human and machine (e.g., in a

radiological setting, X could represent an MRI or some summative characterization of the same picture, such as its complexity; while in a treatment recommendation setting, X could represent both the clinical features of the patients as well as personalization elements that may encode e.g. the preferences of the patient wit respect to different treatment plans): note that X does not need, in general, to be a unique identifier for the cases, but only some relevant characteristics of these latter that are assumed to have a causal effect on both the AI and human ability to formulate a correct judgment. Furthermore, Y is also influenced by I, representing all aspects of the cases that are not used by the machine (e.g., in the previous setting, it could represent any non-digitized information obtained through a patient examination). Finally, D represents the final decision of the human (yet again, whether this decision is correct or wrong), which is obviously influenced by both Y and H only, under the assumptions mentioned above.

The causal approach then aims to quantify the positive and negative dominance effects of the AI intervention Y on the final decision D: following the principles of causal inference, these effects can be computed by *intervening* on Y, that is by fixing the value of Y by means of a *do* operation so as to simulate an intervention performed in a randomized-control trial setting. As in the frequentist approach, we assume that both the final decision D as well as the initial human judgment H are observed: under this assumption there is an unblocked *backdoor path* from Y to D (namely, the undirected path Y-X-H-I-D). Indeed, it can be easily seen from Figure 1 that in the above-mentioned case H is a *collider* and D a descendant of a collider, and both are conditioned upon. Thus, as a consequence, it is not possible to directly estimate the causal effect of the AI intervention on the decision process by means of the raw frequencies in the decision table (as instead is assumed in the frequentist approach) since the above-mentioned unblocked path introduces a spurious and confounding effect. However, the AI intervention can nevertheless be effectively estimated by *adjusting* for X, i.e. conditioning and averaging on the characteristics of the cases at hand. As a consequence, note that collecting such additional information by which to stratify or characterize the cases upon which humans are to make a decision, is necessary to apply the causal approach outlined in this section.

Based on the previous theoretical derivation, we can define the causal version of the automation bias and detrimental algorithmic aversion metrics as (see also Appendix A):

$$\text{Automation Bias} = \frac{A}{1-A}\frac{1-B}{B},$$
$$\text{Detrimental Algorithmic Aversion} = \frac{S}{1-S}\frac{1-T}{T)}$$

where $A = Pr(D = 0|do(Y = 0), H = 1)$, $B = Pr(D = 1|do(Y = 0), H = 1)$, $S = Pr(D = 0|do(Y = 1), H = 0)$ and $T = Pr(D = 1|do(Y = 1), H = 1)$. It is straightforward to observe that both the frequentist and causal approaches are based on the same reliance patterns. Nonetheless, we note that the two approaches can be expected to produce different results in practice. Choice between the two approaches, then, depends on:

(1) Which information is collected in reference to the data production process described in Figure 1;
(2) The dimensionality characteristics of the study and of the collected samples.

In regard to the nature of the collected data, the frequentist approach described in the previous section can be applied whenever one has collected the data needed to compile a decision table as in Table 2, namely whenever one can count the occurrence of the relevant reliance patterns. Whenever the considered study protocol, in addition, allows to collect data at the individual case level and, in particular, data about the relevant characteristics of the considered cases, the causal approach described in this section should also be applied, since it allows to evaluate the causal effects of the AI intervention in a more informative and unbiased manner, as it enables a more fine-grained control of confounders. Taking into account the dimensionality of the collected samples, even though this dimension clearly impacts on the quality of the estimates obtained through both the frequentist and the causal metrics, we note that the adjustment step required in the computation of the causal metrics requires to focus on sub-samples of the data, which may impact the reliability of the obtained estimates when such sub-samples are of too small cardinality. Therefore, selection among the two approaches should take into account the above-mentioned trade-off between sample size and informativeness of the analysis: in any case, the proposed approaches are more complementary than exclusive, and they should be applied together, whenever possible, to get a more comprehensive picture about dominance.

Notwithstanding the above remarks, since both approaches consider odd ratios, we here recall that any estimate of odd ratios should be reported with the corresponding confidence intervals, which for the 95% confidence level can be expressed by applying the following formula [89] $e^{ln(OR)\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$, where $OR$ denotes the value of the odds ratio, expressed in the form $OR = \frac{a/N1}{b/N1}\frac{c/N2}{d/N2}$.

## 5 USER STUDIES

To illustrate the use, and utility, of the metrics proposed above, we show how to apply them in four case studies that we performed in two radiological settings, which regard MRI reading for knee lesion detection and x-ray reading for vertebral fracture detection, one cardiological setting, which regards ECG reading for anomaly detection and classification, and one in oncology, which regards colon endoscopy for lesion detection and characterization. These experimental settings all share the fact that we collected the judgement of the diagnosing physicians *both* before they received the machine's advice, *and* after having received this support, so that a pre-AI (denoted as H, human first judgment) and a post-AI (denoted as D, final decision) comparison could be possible. Figure 2 depict the human-AI interaction protocols of these four case studies in terms of decision processes and the related output artifacts, while Table 3 summarizes their main features.

### 5.1 Methods

The AI intervention encompassed a classification, in all of the cases, as well as a visual explanation in the imaging-based exams (i.e., MRI, x-ray and colonscopy) and a textual explanation in the ECG study. In particular, in 2 studies (MRI and ECG) we decoupled the provisioning of the classification advice and the explanation, while

**Table 3: Summary description of the four case studies.**

| User Study | Num. of physicians | Num. of cases (decisions) | Diagnostic modality | Type of XAI support | Additional info | XAI decoupled from AI |
|---|---|---|---|---|---|---|
| Fracture detection | 7 | 12 (84) | X-ray images | Pixel attribution map | NA | No |
| Colorectal lesion detection | 21 | 504 (10,584) | Endoscopy videos | Region of interest | Average rater confidence | No |
| ECG reading | 44 | 20 (880) | ECG traces | Textual explanations | Case complexity | Yes |
| Knee lesion detection | 12 | 120 (1,440) | MRI images | Pixel attribution map | Self-perceived case complexity | Yes |

in the other 2 studies these two forms of support were always provided together. Finally, in 3 studies (MRI, ECG and colonscopy) we also collected data about the case characteristics, in order to enable the application of the causal metrics in addition to the frequentist ones, while in the remaining study (x-ray) we did not collect such data, due to the small sample of the study.

*5.1.1 The X-ray fracture detection study.* In the orthopedic case study we considered the task of detection of traumatic thoracolumbar fractures from X-ray imaging. We considered a total of 630 vertebral cropped X-rays, which had been collected at the Spine Surgery Centre of the Niguarda General Hospital of Milan (Italy) from 2010 to 2020, from 151 trauma patients over 18 years old, split into 328 no-fracture images (52%) and 302 fracture ones (fractures associated with other conditions than trauma, such as osteoporosis or neoplasms, were excluded from the dataset). For the ground truthing of the training dataset, 3 experienced spine surgeons annotated the 630 images mentioned above, using also Gold Standard CT and MRI images as additional evidence. We involved in the study 7 orthopedists, with varying experience, in the annotation of 12 X-ray images, by means of an online multi-page questionnaire, administered through the Limesurvey platform[5]. For each case, the orthopedists had to first propose a tentative diagnosis (cf, H in Figure 1), after which they were shown the advice of an AI support along with XAI decisional support in the form of a pixel attribution map (rendered in terms of a heat map) and asked to formulate their final decision (cf. D in Figure 1). More details on this study can be found in [14].

*5.1.2 The endoscopic colorectal lesion detection study.* In the oncological case study we considered the task of colorectal lesions detection from endoscopic videos. 21 endoscopists (10 experts, 11 non-experts) from several hospitals (in Austria, Israel, Japan, Portugal and Spain) were involved in the annotation of 504 videos, in 6 batches of 84 videos each. Initially, all the endoscopists evaluated the videos without any advice from the AI, and their initial diagnosis was registered as the first judgment. In particular, endoscopists had to categorize each lesion in five forced-choice options: "Adenoma", "Hyperplastic", "SSL", "Carcinoma", "Uncertain". Their choices were later mapped to the 3-fold output "Adenoma" (corresponding to the choices "Adenoma" or "Carcinoma"), "Non-Adenoma" (for "Hyperplastic" or "SSL" choices), or "Uncertain". Then, in a second separate session conducted after a washout period of at least two weeks after having analyzed the results of the first session, endoscopists were shown the same videos as in the first session, however in this case they were also shown the advice proposed by an AI system , along with the overlay of a green box to highlight the region of suspected lesion detected by the AI support as a form of XAI support, which they used to arrive at a final decision about the case. In both sessions, for each batch and each participant, a different pre-randomized order of presentation of the lesions was prepared, and the endoscopists examined cases at their own pace. For application of the causal metrics, cases were characterized in terms of the average confidence of the involved endoscopists on each case. More details on this study can be found in [80].

*5.1.3 The ECG reading study.* In the cardiological case study we considered the task of reading and classification of heartbeat patterns from electrocardiograms. In this case study, we involved 44 readers, from the Medicine School of the University Hospital of Siena (Italy), with varying level of expertise: 25 less experienced novice readers, and 19 more experienced expert readers. The ECG readers participated in a questionnaire-based study administered through the LimeSurvey platform in which they were asked to annotate 20 ECG cases. The cases were previously extracted from the ECG Wave-Maven repository[6], along with their diagnoses, with the assistance of a resident who was not involved in the questionnaire and selected the cases in order to have a balanced mix of cases with respect to case complexity. As in the other studies, the readers had to first propose a tentative diagnosis, which was recorded. Subsequently, after being shown the advice of the AI support , the respondents could revise their initial diagnosis, providing a second diagnosis, before being shown a textual explanation of the case at hand, as a form of XAI support, and then providing their final diagnosis. For application of the causal metrics, cases were characterized in terms of complexity as reported in the ECG Wave-Maven repository. More details on this study can be found in [15].

*5.1.4 The MRI knee lesion detection study.* In the radiological we considered the task of detection and classification of knee lesions Magnetic Resonance Images (MRI). In this case study, we involved 12 board-certified radiologists by asking them to annotate 120 MRIs extracted from the MRNet dataset, and classify them in terms of lesion presence (either meniscus or anterior cruciate ligament tear) or absence. For each of the cases, the radiologists had then to propose a tentative diagnosis, which was recorded, and then to produce a final classification after that the diagnostic advice of an AI system had been proposed to them. In addition, in half of the cases (i.e., 60) the diagnosis provided by the AI systems was accompanied by an eXplainable AI (XAI) decisional support in the form of a pixel attribution map (rendered in terms of a heat map). This map highlighted the regions of the MRI that were considered to be most relevant by the AI system. For application of the causal metrics,

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

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

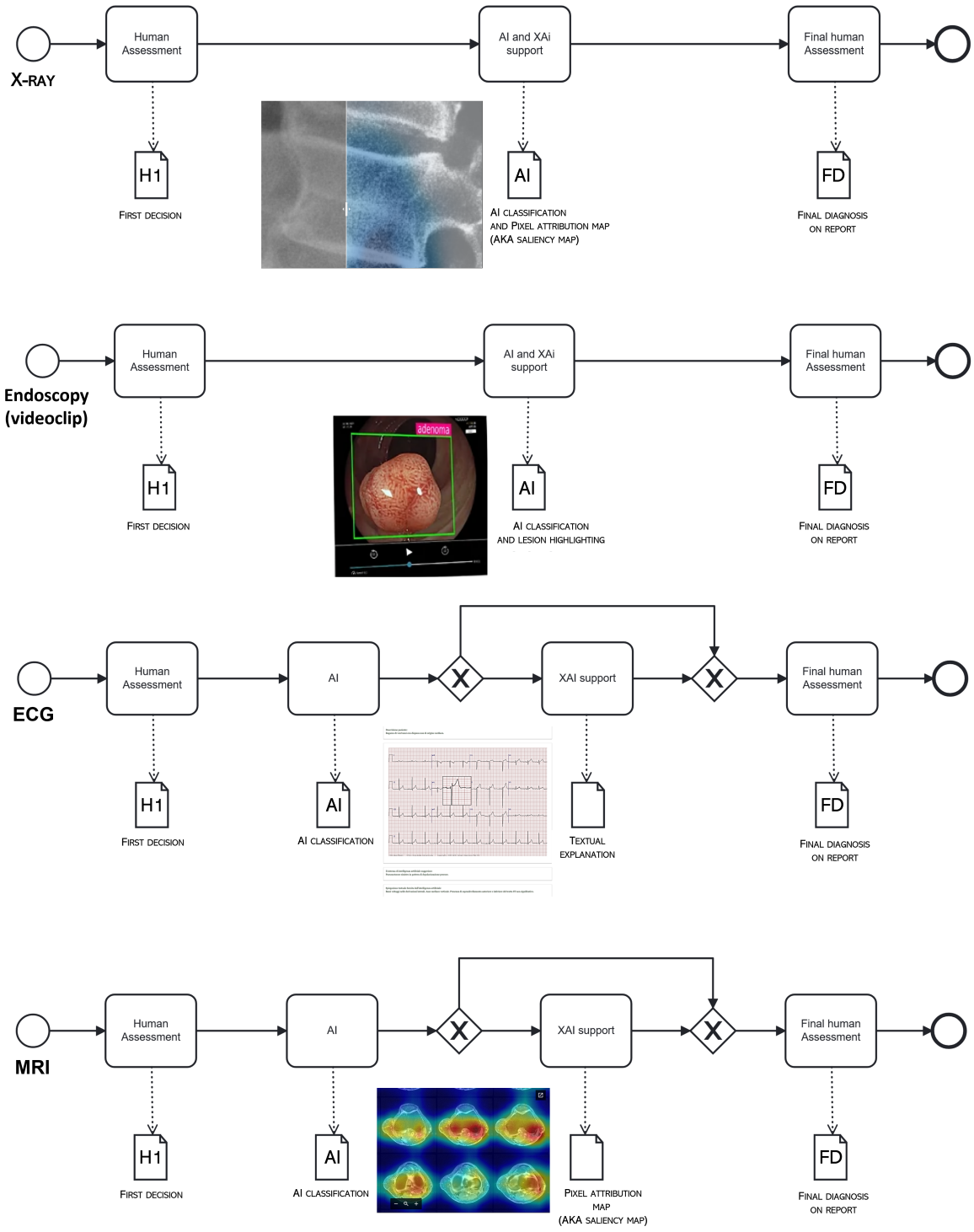


**Figure 2: Human-AI interaction protocols in BPMN notation for each of the 4 case studies. The thumbnails reported at the center of the process schema are extracted from screenshots of the web application used to collect the physicians' decisions.**

cases were characterized in terms of the average perceived complexity, as reported by the the involved radiologists on each case. More details on this study can be found in [15].

## 5.2 Results

For all of the four considered case studies we report the number of occurrences of each reliance pattern in Table 2, in order to compute the frequentist metrics. Furthermore, for all of the considered case studies except the X-ray fracture detection one (for which the required information was not collected), we also collected information about the cases' complexity, to be used to compute the causal metrics. The decision tables for the four case studies are reported in Tables 4, 5, 6 and 7. In all case studies, the *beneficial reliance* pattern was the most frequently observed one (as expected and hoped). More in general, reliance patterns encompassing a correct final decision were more frequently observed, suggesting an overall positive effect of the considered AI interventions.

| H | AI | D | No. Cases | Frequency |
|---|---|---|---|---|
| 0 | 0 | 0 | 82 | .05 |
| 0 | 0 | 1 | 1 | < .01 |
| 0 | 1 | 0 | 238 | .15 |
| 0 | 1 | 1 | 46 | .03 |
| 1 | 0 | 0 | 21 | .01 |
| 1 | 0 | 1 | 202 | .13 |
| 1 | 1 | 0 | 11 | .01 |
| 1 | 1 | 1 | 947 | .61 |

**Table 4: Decision Table of the MRI study. We recall that H denotes the pre-AI human decision; AI the AI advice; D, the definitive (post-AI) decision. Also, 0s stand for wrong decisions, and 1s for the correct ones.**

Indeed, as shown in Figure 4 (which represents in a graphical format the TI metric), in all of the case studies the AI support had a significantly positive effect. According to the interpretation of this metric discussed above, these results confirm the positive effect of the AI interventions that we hinted at above, suggesting a tendency toward overall positive dominance exerted by the AI support,

| H | AI | D | No. Cases | Frequency |
|---|---|---|---|---|
| 0 | 0 | 0 | 78 | .19 |
| 0 | 0 | 1 | 1 | < .01 |
| 0 | 1 | 0 | 45 | .11 |
| 0 | 1 | 1 | 71 | .17 |
| 1 | 0 | 0 | 12 | .03 |
| 1 | 0 | 1 | 35 | .08 |
| 1 | 1 | 0 | 0 | .00 |
| 1 | 1 | 1 | 176 | .42 |

**Table 5: Decision Table of the ECG study. We recall that H denotes the pre-AI human decision; AI, the AI advice; D, the definitive (post-AI) decision. Also, 0s stand for wrong decisions, and 1s for the correct ones.**

| H | AI | D | No. Cases | Frequency |
|---|---|---|---|---|
| 0 | 0 | 0 | 3 | .04 |
| 0 | 0 | 1 | 3 | .04 |
| 0 | 1 | 0 | 3 | .04 |
| 0 | 1 | 1 | 6 | .07 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 22 | .26 |
| 1 | 1 | 0 | 1 | .01 |
| 1 | 1 | 1 | 46 | .55 |

**Table 6: Decision Table of the x-ray study. We recall that H denotes the pre-AI human decision; AI, the AI advice; D, the definitive (post-AI) decision. Also, 0s stand for wrong decisions, and 1s for the correct ones.**

| H | AI | D | No. Cases | Frequency |
|---|---|---|---|---|
| 0 | 0 | 0 | 914 | .09 |
| 0 | 0 | 1 | 228 | .02 |
| 0 | 1 | 0 | 464 | .04 |
| 0 | 1 | 1 | 845 | .08 |
| 1 | 0 | 0 | 313 | .03 |
| 1 | 0 | 1 | 750 | .07 |
| 1 | 1 | 0 | 400 | .04 |
| 1 | 1 | 1 | 6670 | .63 |

**Table 7: Decision Table of the Endoscopy study. We recall that H denotes the pre-AI human decision; AI, the AI advice; D, the definitive (post-AI) decision. Also, 0s stand for wrong decisions, and 1s for the correct ones.**

which nevertheless clearly had a pragmatically useful effect in improving the performance of the medical decision-makers involved in the four studies. Interestingly, the provisioning of XAI support (textual explanations), in addition to the diagnostic advice, lead to an additional significant improvement for the ECG study, while no significant difference was observed (therein in terms of pixel attribution maps) for the MRI study: this latter study was also the one associated with the smallest TI. As a possible reason for this, we notice that the ECG study involved a much higher proportion of students and residents than the MRI case, where only specialists and subspecialists were involved: indeed, as highlighted by recent results in the literature concerned with the evaluation of XAI support [14, 74], less experienced decision-makers may generally have a larger benefit from the introduction of this type of support, while more experienced decision-makers may also be negatively impacted by such systems, possibly due to lower acquaintance with their interpretation. In general these differences suggest that dominance evaluations and TI estimations should be considered spatially and temporally circumscribed: they provide a picture of the effect of a specific AI intervention (an instance of human-AI protocol) within a specific setting and "decision loop" that may change over time, also under the feedback effect of the above intervention (see Figure 3). We will come back to this *decision loop* (which resembles the *task-artifact cycle* theorized by Carroll [21]) later, when we will discuss the framework, also in light of these experimental findings.

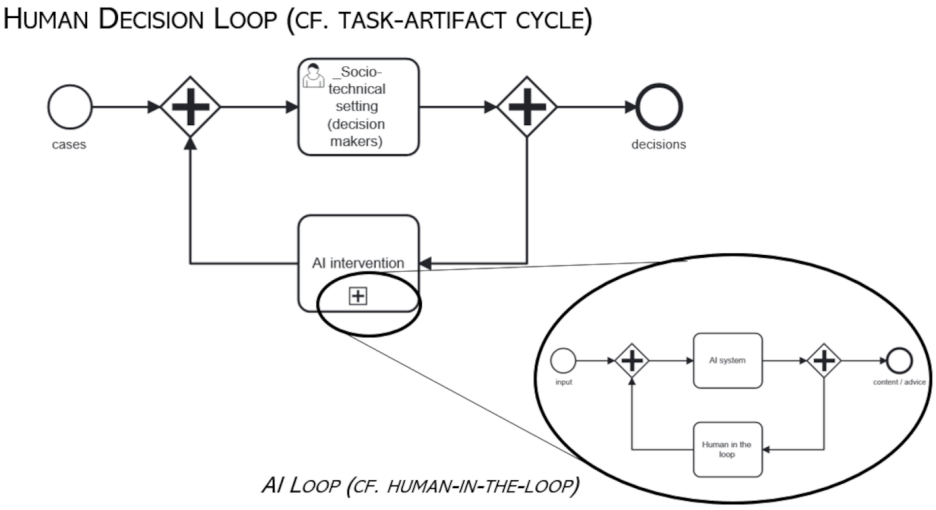


**Figure 3: The decision loop, i.e. the socio-technical system where decisions are made and where a pre-existing natural feedback loop is mediated/augmented by a technological equipment (in this case, an AI intervention). In its turn, this is a system where humans can be involved (so called humans-in-the-loop, but the AI loop!) to improve the system and keep it up to date.**

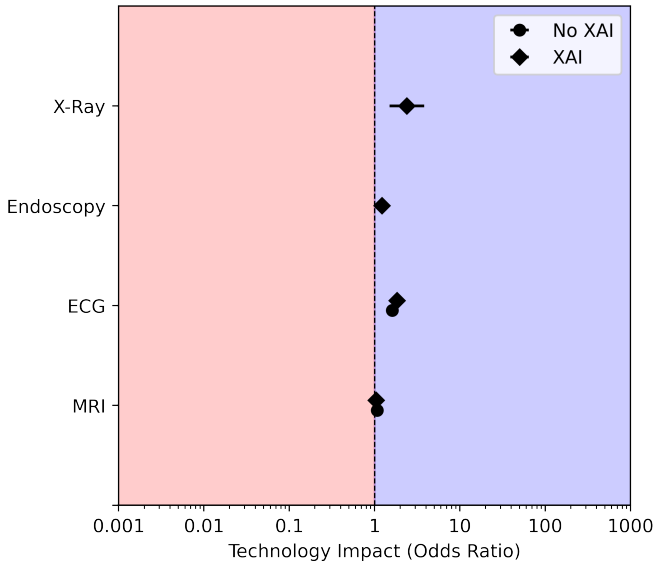


**Figure 4: Technology Impact odds ratios, for the 4 considered case studies. Horizontal lines denote 95% C.I. computed according to the standard formula for odds ratios. The red region denotes an overall negative effect of the AI intervention, while the blue region denotes an overall positive effect.**

Since we detected a significantly positive effect of the AI interventions under study, the frequentist and causal metrics were also applied in order to investigate whether these AI interventions exerted any form of dominance, and the expression of its positive and negative components. The frequentist and causal versions of the Automation Bias metric are reported in Figure 5. In all cases, the observed odds ratios were significantly lower than 1, which can be associated with a greater relative frequency of the *beneficial self-reliance* pattern as compared with the *detrimental over-reliance* pattern, and consequently suggesting the absence of automation bias induced by the AI interventions: indeed, as also further emphasized by means of the causal metrics, the involved study participants tended to rely more on their judgment when the AI provided an erroneous support. Consequently, these results highlight the likely absence of negative dominance and may instead suggest a potential for algorithmic aversion in the studies. Interestingly, according to

both the frequentist and causal analyses, the introduction of XAI support was associated with an higher value of the automation bias odds ratio and, in the case of the causal metrics, this increase was statistically significant. This finding may suggest an increased negative dominance exerted by the addition of the XAI support, which may induce a false sense of understanding in the decision-makers who may then then to over-rely on the AI intervention, even when this latter provided a misguided advice. This rather paradoxical finding has been highlighted also in recent studies concerned with the impact of explanation mechanisms for AI, and has been called in the literature the white-box paradox [37], i.e. the possible increase in decision errors following the introduction of explainable advice. In any case, the reported automation bias odds ratios were significantly smaller than 1, suggesting that the incidence of this white-box paradox phenomenon could be relatively small, at least in the four considered settings.

To better investigate the observed tendency toward algorithmic aversion, the frequentist and causal versions of the Detrimental Algorithmic Aversion odds ratio metric were computed and are reported in Figure 5. As can be noted from the Figures, according to both the frequentist and causal analyses, the participants in the ECG and MRI user studies exhibited a rather skeptical stance when interacting with the AI intervention, having a significant tendency toward algorithmic aversion: indeed, in the MRI case study, as well as in the ECG study when considering the causal analysis, the Detrimental Algorithmic Aversion odds ratios were significantly higher than 1. This result illustrate the likely failure of the considered AI interventions to exert positive dominance, and its associated cognitive effects (i.e. algorithm appreciation), onto the decision-makers involved in the above-mentioned studies. By contrast, in the endoscopy case study the AI intervention exerted some amount of positive dominance, which was confirmed by both the frequentist and causal analyses. These results may seem to confirm the observation according to which one cannot have positive dominance without some portion of the negative one [5] as, indeed, in most of the considered case studies the participants were more often than not likely to reject the advice provided by the AI intervention, both when this latter was wrong but also when it would have been correct to follow it: such behaviour would then explain the observed tendency toward algorithmic aversion, both positive (i.e. the opposite of automation) as well as detrimental. Similarly to the case of the automation bias it can be observed that, according to the causal analysis, the introduction of XAI support lead to an increase in detrimental algorithmic aversion for both the ECG and MRI studies. This finding could be associated with a different manifestation of the white-box paradox, by which explanations (and especially so explanations associated with a decision which is in contrast with one's own) could increase one's self confidence and thus induce in an form of algorithmic aversion. Nonetheless, in contrast with the case of automation bias, the observed increase in detrimental algorithmic aversion due to the introduction of XAI was not statistically significant.

To conclude this section, the results of our analyses allowed us to observe how in the four considered case studies the AI intervention seemed to exert little dominance on the involved participants, with a prevalence of positive, rather than negative, dominance but even more so a tendency towards algorithmic aversion (i.e. the absence of dominance). We were also able to show, by the same means, how the introduction of explanatory hints seemed to reinforce the detrimental interaction with the AI intervention, favoring the emergence of both automation bias and detrimental algorithmic aversion associated behaviours, and thus suggesting the need for further investigation of the so-called white-box paradox. In the following section, we will discuss the implications of our main contributions more widely, that is our proposed methodological framework. Before doing that, however, we want to remark about two limitations of our case studies. First, the main limitation of our study regards the fact that the four studies took place in a laboratory setting, although with real board-certified professionals. This limitation is more universal than common, and we do not mention it to belittle our findings, but rather to join the ranks of those who believe that the stage is ripe for it to become more common for the scientific community to appreciate studies that report results related to the use of certified systems in real world conditions. Nonetheless, we believe this limitation has little effect on our findings, because we are more concerned with the differential effect than we are with absolute performance (hence the adoption of odd ratios). Indeed, whereas absolute accuracy can be affected by experimental conditions, for a sort of Hawthorne or laboratory effect [43, 68], in that real-world performance is generally better, we conjecture that the ratios of AI-human agreement and disagreement would be less prone to changes across different conditions. Second, we remark that the specific insights drawn from our case studies are hardly generalizable to other decision settings: although we believe they are a representative sample of medical diagnostic tasks, these user studies cannot account for the specificities of all possible contexts in which AI support can be applied. Nonetheless, we remark that the application of the proposed metrics can be useful to explore and assess technology dominance in general AI-supported study settings: thus, the case studies were presented with the aim of illustrating which insights could be drawn from the application of our methods, and how these latter can help elucidate the benefits and detriments of introducing AI interventions in a decision scenario.

## 6 DISCUSSION

In this paper we have considered a crucial aspect of human-AI interaction in decision tasks, that is *the influence an AI system exerts on its users, by either changing or confirming their judgment*. We have referred to this influence as *technological dominance*, grounding on the body of research that in the Information Systems field focuses on the design and evaluation of decision support and recommender systems, and that developed a *theory of technological dominance* (TTD) [5, 53, 90]. We adopted this expression, not for a nostalgic retrieval operation of little-trodden lines of research, but to renovate the interest (and currency) about it and to extend it. Indeed, previous contributions in the TTD field have so far mainly focused on the causal factors leading to over-reliance, that is the factors making decision makers more susceptible to technology, and more deferential to its output, both in the short term [53], and in the long one, so paving the way for their deskilling or defensively opportunistic behaviour. These determinants of technology over-dependence include task experience, task complexity, decision aid familiarity, cognitive fit, initial trust, confidence, built trust, and

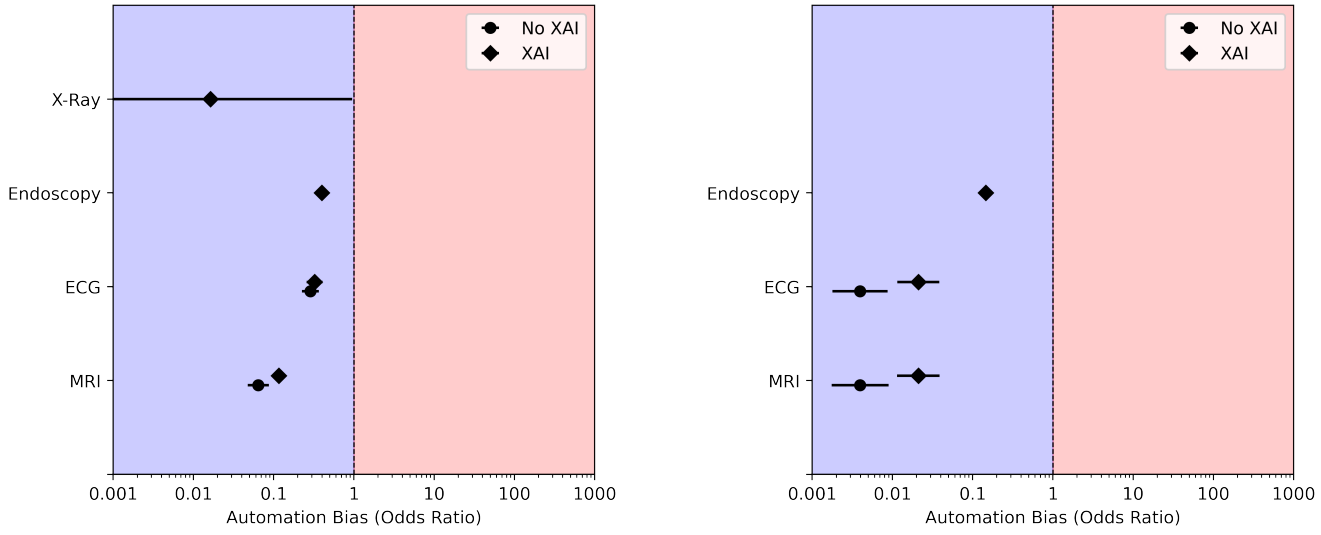


**Figure 5: Automation Bias odds ratios, for the 4 considered case studies: on the left, (a) the frequentist metric; on the right, (b), the causal metric. Horizontal lines denote 95% C.I. computed according to the standard formula for odds ratios. The red region denotes the presence of automation bias due to the AI intervention (i.e. negative dominance), while the blue region denotes the absence of automation bias and the presence of algorithmic aversion.**

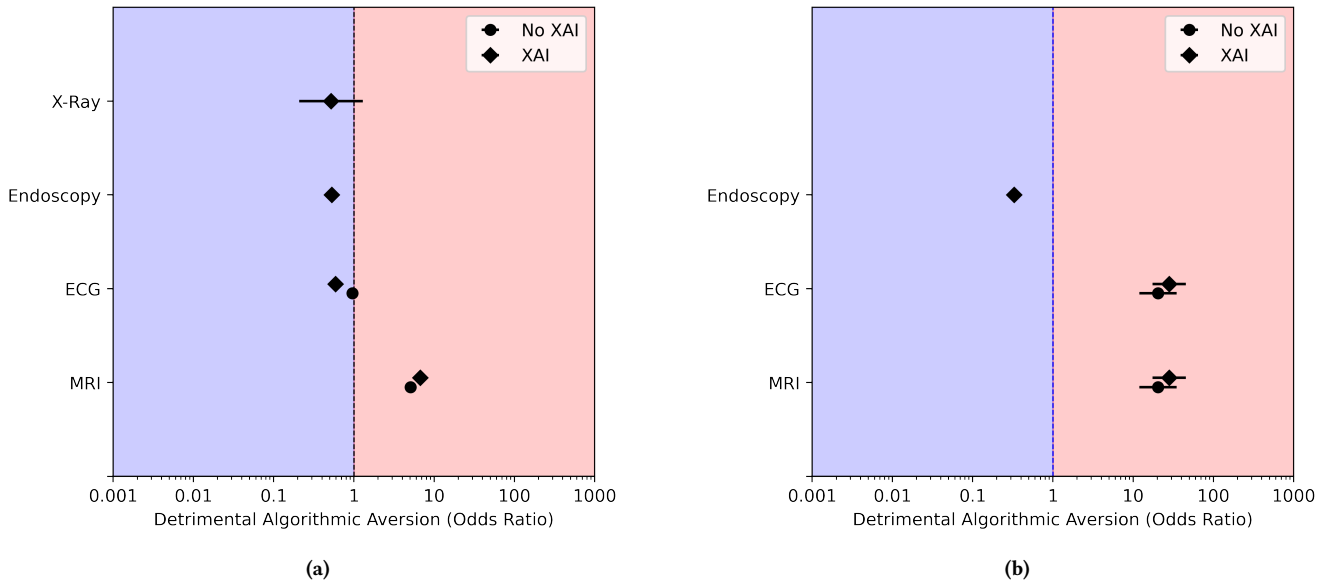


**Figure 6: Detrimental Algorithmic Aversion odds ratios, for the 4 considered case studies: on the left, the frequentist metric; on the right, the causal metric. Horizontal lines denote 95% C.I. computed according to the standard formula for odds ratios. The red region denotes the presence of detrimental algorithmic aversion for the AI intervention, while the blue region denotes the absence of detrimental algorithmic aversion and the presence of algorithm appreciation (i.e. positive dominance).**

explanation, just to mention the ones that recur more frequently in the TTD works. We recognize the importance to research on these factors, and especially to understand which ones can be controlled or changed with socio-technical interventions. At the same time, in this contribution, we purposely have not considered these factors, giving them for granted, as well as their effects; rather, we started from the simple idea that the above interventions *can be* controlled and modulated (by design and in-use adaptations [29]) to mitigate

the effect of technological dominance, and that these interventions can be evaluated (and improved) only if technological dominance is operationalized and is made object of measurement.

Indeed, an AI intervention has many aspects that are under the control of the designer: first of all its output (e.g., categories, probabilities, prioritized lists of alternatives, most similar cases, explanations), but also less-researched design choices such as the order and level of automation (e.g., AI as either a first-opinion supplier or a second-opinion one [15], AI as either an autonomous agent or a case-mining tool [16, 83]), the optimization criteria (e.g., either accuracy or utility [7]), and adaptation to the target population (e.g., either subpsecialists or general practitioners, either novices or experts). Each configuration of these (and other) features is what constitutes a protocol of human-AI interaction, that is a "process schema that stipulates the use of AI tools by competent practitioners to perform a certain task or do a certain job" [15] (see Figures 2). Any protocol instance can be associated to a different potential for influence, reliance by users, and hence dominance after a proper evaluation; on the basis of this evaluation, a specific protocol can be chosen to promote positive effects and mitigate the risk of negative impact. This perspective constitutes the rejection, and also the overcoming, of the usual reasoning according to which the AI is better the more accurate it is (i.e., the fewer mistakes it makes) and for which the best system must be chosen on the basis of its (estimated) performance when this is compared with the average performance of the best human experts or the experts who should use it. The reader will have noticed that in this work, for the four cases reported above, we did not report the basal accuracy of the physicians involved (i.e., the basal accuracy related to the H decision and the frequency of the decision table pattern '1 * *', see Tables 2); nor the accuracy of the AI that was provided to the experts (AI and pattern '* 1 *' Tables 2) nor the final one (D, and pattern '* * 1'). Doing so would have suggested the usual human-machine comparison from which most works (including some ours) do not escape from, when dealing with the adoption of AI systems in critical decision-making domains [19] (such as medicine). Our contribution marks a discontinuity with that evaluative approach, concerned as it is in *differential impact*, rather than improbable performance scores [57, 76] , that is, in the *effect* in terms of either *improvement* and *degradation* of a decision-making team once its members use a decision support (of comparable or slightly greater accuracy) rather than when they do not.

This approach also motivates our reference to the TTD and the use of the term dominance. Although this term has a (purposely) derogatory flavour (which we do not completely disavow), the related concept can also be associated to positive effects and benefit for the overall socio-technical system where technology is adopted: suffice to think of the extent the less experienced users and novices, which are those more prone to technology over-reliance and dominance according to TTD studies [87, 88], could become more integrated in the "decision loop" (see Figure 3) and be involved by virtue of their enhanced condition and "augmentation" by technology. A dominant, i.e., trusted and trustworthy, technology may allow these users to practice during their job while coping with real cases; help them develop analogical reasoning (rather than just flat declarative knowledge of facts, notions, and rules), and propose itself as a sort of transactive memory tapping on a body of past cases and past right decisions [16, 42]. Dominant technologies, in the above sense, can also help the less expert and nonspecialist decision makers achieve (sub)specialist accuracy and, in so doing, democratize [58, 81] access to the expertise that is embedded in the training cases: this is the promise of the *centaur model*, which makes the the human-AI dyad as accurate as the best decision makers when they operate in the best conditions [40]. Thus whether technology dominance has also positive effects largely depends on how human-AI interaction protocols are designed to mitigate loss of skills, inhibition of professional development, opportunistic behaviors (e.g., defensive medicine), and blind faith in the alleged infallibility and objectivity of AI, what has been called *algorithmic authority* [12, 62], or just *algority*.

Our take is that speaking of dominance can promote awareness of the double-edged nature of technology, and above all of our tendency to trust (or distrust) machines thus tipping the balance of power and control when we choose to delegate decisions to external aids. Current research concerning behaviors of reliance usually frames these latter behaviors as expressions of *cognitive biases* (cf. automation bias), and thus as something to be traced back to limits and failures of human individuals. This body of research seems to limit itself to descriptive accounts of these manifestations, thus nurturing a fatalistic approach with respect to the (bad) influence of technology. In light of this body of work, we prefer to propose and adopt a term that relates the potential for bad influence and outcomes to *how* the technology is designed and deployed (i.e., dominance), thus holding designers and programmers responsible, rather than to how our cognition works and to its alleged shortcomings (i.e., biases), which nevertheless can be traced back to comprehensible (and often effective) least-effort heuristics [38, 84].

Thus, in this work we did not try to associate different causes of dominance with its observable effects, nor with possible courses of action for AI design improvement, although we believe this goal should attract further research in future. Rather, we aimed to shed light on a preliminary step that makes the above goals possible in the context of empirical research, such as observational studies or research trials: the quantitative estimation of dominance and the assessment of the "size" of its positive and negative components. Measuring dominance is one step, we believe a necessary one, of a longer process of recognition and responsibility taking, by both designers and users, about the consequences of automation in decision tasks, which also entails the conception and implementation of prevention or mitigation measures. As said above, this can be done by comparing different interaction protocols (and interventions) that differ for modifiable features (such as the provision of explanations, as the XAI research field dictates), and by implementing the protocols that are associated with higher estimates of effectiveness (i.e., *positive technological impact*).

To this aim, we introduced simple metrics to estimate dominance, in a manner equivalent to how one estimates the accuracy of a system from its performance on a test dataset (i.e., in terms of either out-of-sample or out-of-distribution prediction), on the basis of a kind of extended confusion matrix, which we call *decision table*. In particular, we distinguished two different approaches to measure dominance, in general: first, a frequentist approach, which is directly based on the above-mentioned confusion matrices and can be applied to any study in which such data about the interaction of

a human decision-maker and an AI system has been collected; second, a causal inference approach, which adds to and complements the frequentist one by providing more reliable and generalizable estimates of the causal effect of introducing an AI intervention in a decision-making setting, by properly considering some characteristics of the cases at hand. In addition to this, we also proposed some specific dimensions of dominance: automation bias and detrimental algorithmic appreciation; and its opposite, algorithm appreciation, giving formal definitions and formulas according to the above general statistical approaches. We remark that the applicability of the proposed approach is limited to binary right-or-wrong decisions: future work should extend our proposal to account for ordinal or numeric judgments and predictions, as well as for the varying levels of confidence reported by either the human or the AI.

As we commented above, the defined metrics can be easily applied in longitudinal studies, since they require an explicit distinction between pre-AI and post-AI human assessments, while their application in cross-sectional studies may be more difficult: this could be achieved either by adopting study designs that enable the appropriate matching and comparison of unaided and aided groups (e.g. pseudo-longitudinal studies [6] or matched repeated cross-sectional studies [59]); or by considering cross-sectional studies in which a full interaction history with the AI intervention is considered as a single time point (e.g. by comparing two different AI interventions). While this is certainly a limitation of the proposed framework, we believe that this is intrinsic in the assessment of dominance, as this latter dimension emphasizes the role of an AI intervention in affecting human decision-making and fostering decision changes, and these changes necessarily occur over time.

A similar limitation relates to the application of the proposed metrics to so-called AI-first decision settings, where, contrarily to the scenario assumed in this article, the AI intervention support is given to the human users at the same time as the case data and information. In these settings, obviously, the initial, unaided human decisions (i.e. H in Figure 1 and Table 2) is not observed and cannot be directly used for the computation of our metrics. Nonetheless, their value can be estimated (or, better yet, bounded) by using standard approaches for dealing with latent variables that have been proposed in the statistics literature [10], and in particular by using interval estimation methods [33]. In this case, the initial decision can be simulated, by separately calculating the metrics under both settings of the unobserved initial decision (i.e. by alternately assuming that this latter was correct/wrong), for each case. In doing so, lower and upper bounds on the values of the metrics can be computed. We plan to better investigate this approach in future work.

Finally, we want to remark on a limitation of our framework, which regards the intertwining of *personalization* and *repeated* (or, *sequential*) interaction. As we previously described, our framework allows to capture some form of personalization information, when these can be understood to be part of the X variable in Figure 1. This requires, first of all, that the user (or, rather, their characteristics of interest) can be represented in a digitized form and, most importantly, that the characteristics of interest should be static (they do not change over the course of the interaction between the user and the AI system) and already defined at the beginning of the interaction. The reason behind these restrictions is that our proposed framework assumes that, for each case of interest, the interaction between the AI system and the user is one-shot: the interaction begins when the case is presented and ends when the human provides a final decision for it. Obviously, there could be repeated interactions between the AI system and the users (trivially, the set of cases that is used for evaluating the impact of AI support on the decision-making is obtained as the result of such a repeated interaction); however, any personalization information (in the sense above) and also the AI system itself should not change within this repeated interaction. These assumptions are restrictive yet sufficiently encompassing: in the medical domain, for example, they would allow having different profiles of doctors depending on their skill (e.g., as measured on a standardized set of previous cases), or their diagnostic attitude in terms of sensitivity and specificity and treatment preferences. At the same time, it is clear that not all possible interaction scenarios may be reduced to the above-mentioned instances: for example, any setting in which there is a gradual fine-tuning of the AI system, or when the AI system tries to dynamically infer some characteristic of the present user *during the current interaction*, is out of the scope of, and hence cannot be studied through, the framework that we propose here. Thus, we believe that future work should also be aimed at exploring how our proposed approach could be generalized to encompass these more general interaction scenarios. A promising direction would be to extend the proposed causal model (see Figure 1) to allow causal dependencies between the final human decision (node D) and the AI support (node Y); doing so would enable the modeling of temporal and sequential dynamics in the interaction between the AI and the user, by employing extensions of causal diagram that explicitly model the time dimension (e.g. dynamic causal networks [8]).

## 7 CONCLUSION

As said above, our contribution extends the research conducted under the tenets of the TTD. Indeed, while previous studies on technology dominance have focused on the determinants of the influence of technological systems on our judgment and discretion, this is the first work adopting TTD to give an analytical interpretation of phenomena that have different names in the literature: namely, automation bias, automation aversion, automation appreciation, that is when AI misleads us, when we reject its advice even when it is right, and when we follow it (and beneficially so), respectively. In this sense, the framework that we propose allows to distinguish the positive component of dominance, which is related to improvements in decision accuracy [20], from the negative ones, which are related to automation bias and detrimental algorithmic aversion, mainly (and their long-term effects [17, 67]).

As we discussed in Section 2, our framework shares some characteristics with, and builds upon, the recent contributions of Schemmer et al. [82] and Reverberi et al. [80]. In particular, compared to the work of Schemmer, we extend it in regard to what they first called "reliance patterns" with our notion of *decision table*, and a first endevour to link these patterns with the main related cognitive biases and effects. Furthermore, in contrast with the metrics proposed by Schemmer et al., our metrics are based on a principled statistical framework that allows for the quantification of uncertainty (via the computation of odds ratio metrics and their

associated confidence intervals), as well as of the causal influence of an AI intervention. In regard to the proposal by Reverberi et al. [80], we recall that, as discussed in Section 4, the TI indicator proposed in this work directly corresponds to the inverse of the "effect on diagnostic accuracy" that is defined in [80]. As discussed before, our approach provides a closer correspondence with the usual definition and interpretation of odds ratio for evaluating the effectiveness of interventions in empirical research. Furthermore, we note that even though the "effectiveness" (resp. "safety") and Detrimental Algorithmic Aversion (resp. Automation Bias) odds ratio are related to the notion of algorithm appreciation (resp. automation bias), the proposed metrics provide a quantification of dominance at a finer level of granularity since they consider the initial judgment formulated by the human decision makers together with the corresponding final decision (see Appendix A). Thus, we believe that our proposal provides a more comprehensive way to assess the effect of AI on decision-making in that it also allows the interested practitioners to precisely decompose the positive and negative dimensions of dominance.

In addition to these contributions, we also distinguished between the effect of the AI's advice and the effect of AI explanations, thus introducing a way to quantify the so called "white-box paradox" that some studies have recently begun to observe [15, 37], that is the effect occurring when good (i.e., persuasive) explanations leads to automation bias, and therefore users discard their own correct judgment to follow the wrong AI advice. A taxonomy of reliance patterns, the notion of decision table, and data visualizations to "see" AI influence and impact (i.e., odds ratio visualizations) are the other contributions of this work, which have been proposed to get a picture of the potential influence of AI-based decision aids in decision-making, and were illustrated in four studies that involved physicians from different specialities in simulated but realistic diagnostic tasks.

We hope that our framework could serve as a basis for future work in the evaluation of the effect of the introduction of AI support in real-world decision-making settings, for its capability of providing the interested researchers and practitioners with metrics and tools by which they can evaluate the impact and dominance of AI, and therefore adjust the "parameters" of human-AI interaction protocols to improve hybrid decision-making [1]. In this sense, we do not intend the proposed metrics as possible subject of optimization, or incentive structures, because they too could be subject to Goodhart effects [41] and performative feedback loops [48]: instead, we propose these metrics and definitions for *technovigilance* initiatives [18, 30] and the continuous monitoring of the usability (that is effectiveness, efficiency, and satisfaction) of AI-based decision supported systems deployed and adopted in specific work settings. On a pratical level, these assessment initiatives require that a certain proportion of randomly selected cases are given to the human decision makers *before* they can access the AI advice, so that pre-AI decisions can be recorded, and then compared with the definitive post-AI decisions (H and D, respectively), and the metrics we propose be applied (once a ground truth reference has been established also for the prospective cases).

To conclude: We agree with Deming that "it is wrong to suppose that if you can't measure it, you can't manage it" [26]; nevertheless, we believe that implementing an AI-based decision support system without looking at the effects that this can bring to the human "decision loop" into which it is socially embedded would be as irresponsible as driving blindfolded. The metrics of technological impact and dominance that we propose in this work are a contribution in the direction of *responsible AI* [28], not only in data collection, model definition and system design, but also in use and adoption, that is "in the wild" [56].

## REFERENCES


[1] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 08 (2020), 18–28.

[2] Saar Alon-Barkat and Madalina Busuioc. 2023. Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory* 33, 1 (2023), 153–169.

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

[4] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine* 25, 6 (2019), 954–961.

[5] Vicky Arnold and Steve G Sutton. 1998. The theory of technology dominance: Understanding the impact of intelligent decision aids on decision maker's judgments. *Advances in accounting behavioral research* 1, 3 (1998), 175–194.

[6] Alapan Bandyopadhyay and Abhijit Mukherjee. 2021. Pseudo-longitudinal research design: a valuable epidemiological tool in resource-poor settings. *Rural and Remote Health* 21, 4 (2021), 6977.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? Optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. AAAI, Washington, DC (USA), 11405–11414.

[8] Gilles Blondel, Marta Arias, and Ricard Gavaldà. 2017. Identifiability and transportability in dynamic causal networks. *International journal of data science and analytics* 3, 2 (2017), 131–147.

[9] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

[10] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. 2021. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics* 49, 5 (2021), 2885–2915.

[11] Jacques Bouaud, Jean-Philippe Spano, Jean-Pierre Lefranc, Isabelle Cojean-Zelek, Brigitte Blaszka-Jaulerry, Laurent Zelek, Axel Durieux, Christophe Tournigand, Alexandra Rousseau, Pierre-Yves Vandenbussche, and Brigitte Séroussi. 2015. Physicians' Attitudes Towards the Advice of a Guideline-Based Decision Support System: A Case Study With OncoDoc2 in the Management of Breast Cancer Patients. *Studies in Health Technology and Informatics* 216 (2015), 264–269.

[12] Federico Cabitza. 2021. *Cobra AI: Exploring Some Unintended Consequences.* MIT Press, Cambridge, MA (USA), 87.

[13] Federico Cabitza, Andrea Campagner, and Edoardo Datteri. 2021. To err is (only) human. Reflections on how to move from accuracy to trust for medical AI. In *Exploring Innovation in a Digital World.* Springer, Cham (Switzerland), 36–49.

[14] Federico Cabitza, Andrea Campagner, Lorenzo Famiglini, Enrico Gallazzi, and Giovanni Andrea La Maida. 2022. Color Shadows (Part I): Exploratory Usability Evaluation of Activation Maps in Radiological Machine Learning. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction.* Springer, Cham, Switzerland, 31–50.

[15] Federico Cabitza, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa. 2023. Rams, Hounds and White Boxes: Investigating Human-AI Collaboration Protocols in Medical Diagnosis. *Artificial Intelligence in Medicine* Forthcoming (2023).

[16] Federico Cabitza, Andrea Campagner, and Carla Simone. 2021. The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies* 155 (2021), 102696.

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

[18] Federico Cabitza and Jean-David Zeitoun. 2019. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Annals of translational medicine* 7, 8 (2019).

[19] Francisco Maria Calisto, Nuno Nunes, and Jacinto C Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies* 168 (2022), 102922.

[20] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. *Artificial Intelligence in Medicine* 127 (2022), 102285.

[21] John M Carroll, Wendy A Kellogg, and Mary Beth Rosson. 1991. *The task-artifact cycle.* Cambridge University Press, Cambridge (UK), 74–102.

[22] Enrico Coiera. 2019. Assessing technology success and failure using information value chain theory. *Stud Health Technol Inform* 263 (2019), 35–48.

[23] Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. In *AIAA 1st Intelligent Systems Technical Conference.* American Institute of Aeronautics and Astronautics, Reston, VA (USA).

[24] Jacob Davis, Andrew Atchley, Hannah Smitherman, Hailey Simon, and Nathan Tenhundfeld. 2020. Measuring Automation Bias and Complacency in an X-Ray Screening Task. In *2020 Systems and Information Engineering Design Symposium (SIEDS).* IEEE, Charlottesville, VA, USA, 1–5. https://doi.org/10.1109/SIEDS49339.2020.9106670

[25] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376638

[26] W Edwards Deming. 2018. *The new economics for industry, government, education.* MIT press, Cambridge, MA (USA).

[27] Berkeley J Dietvorst, Joseph Simmons, and Cade Massey. 2014. Understanding algorithm aversion: forecasters erroneously avoid algorithms after seeing them err. In *Academy of Management Proceedings.* Academy of Management, Briarcliff Manor, NY (USA), 12227.

[28] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way.* Springer Nature, Berlin (Germany).

[29] Alan Dix. 2007. Designing for appropriation. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21.* BCS Learning & Development Ltd., Swinton (UK), 1–4.

[30] Mary Dixon-Woods, Sabi Redwood, Myles Leslie, Joel Minion, Graham P Martin, and Jamie J Coleman. 2013. Improving quality and safety of care using "technovigilance": an ethnographic case study of secondary use of data from an electronic prescribing and decision support system. *The Milbank Quarterly* 91, 3 (2013), 424–454.

[31] Naomi J. Dunn, Thomas A. Dingus, Susan Soccolich, and William J. Horrey. 2021. Investigating the impact of driving automation systems on distracted driving behaviors. *Accident Analysis & Prevention* 156 (June 2021), 106152. https://doi.org/10.1016/j.aap.2021.106152

[32] Joann G Elmore and Christoph I Lee. 2022. Artificial Intelligence in Medical Imaging—Learning From Past Mistakes in Mammography. *JAMA Health Forum* 3, 2 (2022), e215207.

[33] Carl F Falk and Jeremy C Biesanz. 2015. Inference and interval estimation methods for indirect effects with latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal* 22, 1 (2015), 24–38.

[34] Brian J Fogg. 1998. Captology: the study of computers as persuasive technologies. In *CHI 98 Conference Summary on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY (USA), 385.

[35] Brian Jeffrey Fogg, Gregory Cueller, and David Danielson. 2007. Motivating, influencing, and persuading users: An introduction to captology. In *The human-computer interaction handbook.* CRC press, Boca Raton, FL (USA), 159–172.

[36] Charles P Friedman. 2009. A "fundamental theorem" of biomedical informatics. *Journal of the American Medical Informatics Association* 16, 2 (2009), 169–170.

[37] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[38] Gerd Gigerenzer and Henry Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science* 1, 1 (2009), 107–143.

[39] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics* 83, 5 (2014), 368–375.

[40] Ian M Goldstein, Julie Lawrence, and Adam S Miner. 2017. Human-machine collaboration in cancer and beyond: The centaur care model. *JAMA oncology* 3, 10 (2017), 1303–1304.

[41] Charles AE Goodhart. 1984. Problems of monetary management: the UK experience. In *Monetary theory and practice.* Palgrave, London (UK), 91–121.

[42] Pranav Gupta and Anita Williams Woolley. 2021. Articulating the role of artificial intelligence in collective intelligence: A transactive systems framework. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 65. SAGE Publications, Los Angeles, CA (USA), 670–674.

[43] David Gur, Andriy I Bandos, Cathy S Cohen, Christiane M Hakim, Lara A Hardesty, Marie A Ganott, Ronald L Perrin, William R Poller, Ratan Shah, Jules H Sumkin, et al. 2008. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 249, 1 (2008), 47.

[44] Joseph Y Halpern. 2016. *Actual causality.* MIT Press, Cambridge, MA (USA).

[45] Clark Hampton. 2005. Determinants of reliance: An empirical test of the theory of technology dominance. *International Journal of Accounting Information Systems* 6, 4 (2005), 217–240.

[46] Clare Harries, Ilan Yaniv, and Nigel Harvey. 2004. Combining advice: The weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making* 17, 5 (2004), 333–348.

[47] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes* 70, 2 (1997), 117–133.

[48] Mireille Hildebrandt. 2022. The Issue of Proxies and Choice Architectures. Why EU law matters for recommender systems. *Frontiers in Artificial Intelligence* 5 (2022), 73.

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

[50] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[51] Weiwei Huo, Guanghui Zheng, Jiaqi Yan, Le Sun, and Liuyi Han. 2022. Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. *Computers in Human Behavior* 132 (2022), 107253.

[52] Makoto Itoh and Kenji Tanaka. 2000. Mathematical modeling of trust in automation: Trust, distrust, and mistrust. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 44. SAGE Publications, Los Angeles, CA (USA), 9–12.

[53] Matthew L Jensen, Paul Benjamin Lowry, Jude K Burgoon, and Jay F Nunamaker. 2010. Technology dominance in complex decision making: The case of aided credibility assessment. *Journal of Management Information Systems* 27, 1 (2010), 175–202.

[54] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *Twenty-Eighth European Conference on Information Systems (ECIS2020).* AIS, Atlanta, GA (USA), 1–16.

[55] Daniel Kahneman, Olivier Sibony, and CR Sunstein. 2022. *Noise.* HarperCollins UK, Glasgow (Scotland).

[56] K Katsikopoulos, O Simsek, Marcus Buckmann, and Gerd Gigerenzer. 2020. *Classification in the Wild.* MIT Press, Cambridge, MA (USA).

[57] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine* 17, 1 (2019), 1–9.

[58] Yasuyuki Kobayashi, Maki Ishibashi, and Hitomi Kobayashi. 2019. How will "democratization of artificial intelligence" change the future of radiologists? *Japanese journal of radiology* 37, 1 (2019), 9–14.

[59] Matthew J Lebo and Christopher Weber. 2015. An effective approach to the repeated cross-sectional design. *American Journal of Political Science* 59, 1 (2015), 242–258.

[60] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, and Alastair K Denniston. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *bmj* 370 (2020), 1364—-1374.

[61] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[62] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY (USA), 1057–1062.

[63] David Lyell and Enrico Coiera. 2016. Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association* 24 (08 2016), ocw105.

[64] David Lyell and Enrico Coiera. 2017. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* 24, 2 (2017), 423–431.

[65] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. arXiv:2301.05809

[66] John J Masselli, Robert C Ricketts, Vicky Arnold, and Steve G Sutton. 2002. The Impact of Embedded Intelligent Agents on Tax-Reporting Decisions. *Journal of the American Taxation Association* 24, 2 (2002), 60–78.

[67] Anne-Sophie Mayer, Franz Strich, and Marina Fiedler. 2020. Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive* 19, 4 (2020).

[68] Rob McCarney, James Warner, Steve Iliffe, Robbert Van Haselen, Mark Griffin, and Peter Fisher. 2007. The Hawthorne Effect: a randomised, controlled trial. *BMC medical research methodology* 7, 1 (2007), 1–8.

[69] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[70] Stephanie M. Merritt, Alicia Ako-Brew, William J. Bryant, Amy Staley, Michael McKenna, Austin Leone, and Lei Shirase. 2019. Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology* 10 (2019).

[71] Kathleen L Mosier and Linda J. Skitka. 1996. Human Decision Makers and Automated Decision Aids: Made for Each Other? In *Automation and Human Performance: Theory and Applications*, Raja Parasuraman and Mustapha Mouloua (Eds.). CRC Press, Boca Raton, Chapter 10, 201–220.

[72] Tracy Noga and Vicky Arnold. 2002. Do tax decision support systems affect the accuracy of tax compliance decisions? *International Journal of Accounting Information Systems* 3, 3 (2002), 125–144.

[73] Dilek Önkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22, 4 (2009), 390–409.

[74] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. *Advances in Neural Information Processing Systems* 34 (2021), 610–623.

[75] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (2000), 286–297.

[76] Seong Ho Park and Kyunghwa Han. 2018. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 286, 3 (2018), 800–809.

[77] Judea Pearl. 2012. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Arlington, VA (USA), 3–11.

[78] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books, New York, NY (USA).

[79] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6 (March 2017), 691–702. https://doi.org/10.1002/for.2464

[80] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports* 12, 1 (2022), 1–10.

[81] Giovanni Rubeis, Keerthi Dubbala, and Ingrid Metzler. 2022. "Democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Frontiers in Genetics* 13 (2022).

[82] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. In *CHI Conference on Human Factors in Computing Systems (CHI '22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*. Association for Computing Machinery, New York, NY (USA), 10.

[83] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[84] Linda J Skitka, Kathleen Mosier, and Mark D Burdick. 2000. Accountability and automation bias. *International Journal of Human-Computer Studies* 52, 4 (2000), 701–717.

[85] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.

[86] Daniel Susser and Vincent Grimaldi. 2021. Measuring Automated Influence: Between Empirical Evidence and Ethical Values. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 242–253.

[87] Steve G Sutton, Vicky Arnold, and Matthew Holt. 2018. How much automation is too much? Keeping the human relevant in knowledge work. *Journal of emerging technologies in accounting* 15, 2 (2018), 15–25.

[88] Steve G Sutton, Vicky Arnold, and Matthew Holt. 2022. An Extension of the Theory of Technology Dominance: Understanding the Underlying Nature, Causes and Effects.

[89] Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry* 19, 3 (2010), 227.

[90] Anis Triki and Martin M Weisner. 2014. Lessons from the literature on the theory of technology dominance: Possibilities for an extended research framework. *Journal of Emerging Technologies in Accounting* 11, 1 (2014), 41–69.

[91] Bradley Edward Williams. 2020. *The Role of Complexity Within Intelligent Decision Aids on User Reliance: An Extension of the Theory of Technology Dominance*. Ph.D. Dissertation. The University of North Carolina at Charlotte.

# A TECHNICAL APPENDIX

In Section 4.1 we provided intuitive formulas for the AIE, AIN, CE and CN terms. In this appendix, we show how these definitions can be given a formulation based on decision tables. Clearly, the AIE and AIN terms can be easily be defined in terms of the patterns appearing in a decision table. Indeed:

$$AIE = \frac{\text{dr} + \text{dsr} + \text{dor} + \text{dur}}{N}$$

$$AIN = \frac{\text{bur} + \text{bor} + \text{bsr} + \text{br}}{N}$$

where $dr$ = detrimental reliance, $dsr$ = detrimental self-reliance, $dor$ = detrimental over-reliance, $dur$ = detrimental under-reliance, $bur$ = beneficial under-reliance, $bor$ = beneficial over-reliance, $bsr$ = beneficial self-reliance and $br$ = beneficial reliance. More remarkably, also the CE and CN terms can be defined in terms of decision table patterns, even though their derivation is less immediate. To understand this derivation, note that a decision table enumerates all possible combinations of the H, Y and D random variables (see Table 2 and Figure 1): therefore, the frequencies of these combinations define a joint distribution for the three variables mentioned above. CE and CN refer, respectively, to the cases $H = 0$ and $H = 1$, i.e. to the marginal on variable $H$ of the above-mentioned joint distribution. Therefore, CE and CN can be computed by marginalization, as follows:

$$CE = \frac{\text{dr} + \text{bur} + \text{dsr} + \text{bor}}{N}$$

$$CN = \frac{\text{dor} + \text{bsr} + \text{dur} + \text{br}}{N}$$

In Section 4.2, the causal versions of the automation bias and detrimental algorithmic aversion metrics have been defined as:

$$\text{Automation Bias} = \frac{A}{1-A}\frac{1-B}{B},$$

$$\text{Detrimental Algorithmic Aversion} = \frac{S}{1-S}\frac{1-T}{T)},$$

where $A = Pr(D = 0|do(Y = 0), H = 1)$, $B = Pr(D = 1|do(Y = 0), H = 1)$, $S = Pr(D = 0|do(Y = 1), H = 0)$ and $T = Pr(D = 1|do(Y = 1), H = 1)$. In the above formulas, the $do(Y = k)$ statements refer to the intervention that fixes the value of the random variable $Y$ to $k$ (where $k \in \{0, 1\}$). However, since the involved probabilities involve conditioning on an intervention event, their values cannot be directly be computed by the conditional probability formula. Thus, based on do-calculus [77], the above probabilities can be estimated through the adjustment formula as follows:

$$Pr(D = d|do(Y = y), H = h) = \sum_c P(d, y, h, c)Z(c, h),$$

where $P(d, y, h, c) = Pr(D = d|Y = y, H = h, X = c)$ and $Z(c, h) = Pr(X = c|H = h)$. Intuitively, the above formula states that, in order to compute the effect of fixing the AI support through an intervention, one should marginalize over all possible values of the confounding variable $X$ conditioned over the possible initial human judgments (see also Figure 1).

In Section 7, we briefly hinted at the relationship between our approach and the previous work in [80]. To illustrate this claim, let us take the case of the "effectiveness" as an example, this latter is defined in [80] as shown in Table 1 and can be expressed in the terminology of our framework as:

$$\frac{\text{bor} + \text{br}}{\text{dsr} + \text{dur}}\frac{\text{dsr} + \text{bor}}{\text{dur} + \text{br}}.$$

Doing so (while clearly taking into account the algorithm appreciation) conflates this effect also with what we denoted as confirmation bias, which is not directly related to positive dominance. By contrast, the (inverse of) proposed Detrimental Algorithm Aversion odds ratio correctly accounts only for the patterns associated with positive dominance.