AI Persuasion, Bayesian Attribution, and Career Concerns of Doctors*

Hanzhe Li Jin Li Ye Luo Xiaowei Zhang

September 27, 2024

Abstract

This paper examines how AI persuades doctors when their diagnoses differ. Disagreements arise from two sources: attention differences, which are objective and play a complementary role to the doctor, and comprehension differences, which are subjective and act as substitutes. AI's interpretability influences how doctors attribute these sources and their willingness to change their minds. Surprisingly, uninterpretable AI can be more persuasive by allowing doctors to partially attribute disagreements to attention differences. This effect is stronger when doctors have low abnormality detection skills. Additionally, uninterpretable AI can improve diagnostic accuracy when doctors have career concerns.

Keywords: AI, persuasion, attribution, interpretability, diagnosis, career concerns

JEL classifications: D02, D83, I10, I30, M16

*Contact information: Hanzhe Li, University of Hong Kong, hanzhe.li@connect.hku.hk; Jin Li, University of Hong Kong, jli1@hku.hk; Ye Luo, University of Hong Kong, kurtluo@hku.hk; Xiaowei Zhang, Hong Kong University of Science and Technology, xiaoweiz@ust.hk. We thank Sofia Bapna, Dan Barron, Yeon-Koo Che, Yi Chen, Kim-Sau Chung, Bo Cowgill, Wouter Dessein, Jeff Ely, Florian Englmaier, Matthias Fahn, Ricard Gil, Luis Garicano, Jie Gong, Yao Huang, Carl Heese, Nan Jia, Rongzhu Ke, Shaowei Ke, Peter Klibanoff, John Kolstald, Wing Tung Lam, Benson Leung, Fei Li, Yalun Li, Ting Liu, Susan F. Lu, Zhuoran Lu, Xueming Luo, Yele Ma, Vincent Meisner, Masaki Miyashita, Huifeng Su, Junze Sun, Paul Seabright, Scott Schaefer, Wing Suen, Balazs Szentes, Feng Tian, Brian Viard, Brian Wu, Yanhui Wu, Chen Zhao, and participants at Columbia/Wharton MAD conference, HKEA biennial conference, OESS, POMS-HK conference for their helpful conversations and comments.

1 Introduction

> *People should stop training radiologists now. It's just completely obvious within five years deep learning is going to do better than radiologists.*
>
> - Geoffrey Hinton, *2016*

> *The big claims about AI assume that if something is possible in theory, then it will happen in practice. That is a big leap.*
>
> - Peter Cappelli and Valery Yakubovich, *2024*

Although AI has the potential for transformative impact (Agrawal et al., 2018a; Brynjolfsson et al., 2021; Agrawal et al., 2022; Grossmann et al., 2023), the effectiveness of human-AI collaboration is widely recognized as a crucial factor in AI's efficacy (Beede et al., 2020; Gruber et al., 2020; Mullainathan and Obermeyer, 2021; DeStefano et al., 2022; Chen and Chan, 2023; Dell'Acqua et al., 2023a,b; Otis et al., 2023; Cao et al., 2024; Vanneste and Puranam, 2024; Wang et al., 2024). Several studies document that humans frequently resist AI suggestions for various reasons, indicating an AI-aversion problem (e.g., Dietvorst et al., 2015; Longoni et al., 2019; Luo et al., 2019; Kawaguchi, 2021; Tong et al., 2021; Liu et al., 2023; Kim et al., 2024; Yin et al., 2024; see Burton et al., 2020, Mahmud et al., 2022, and De Freitas et al., 2023 for surveys).

In the medical field, in particular, AI adoption appears slower than expected (Dranove and Garthwaite, 2022). Doctors often do not incorporate AI suggestions into their diagnoses (Clement et al., 2021; Jussupow et al., 2021; Lebovitz et al., 2022; Agarwal et al., 2023; Yu et al., 2024). One explanation for this is that doctors are career-concerned (Arkes et al., 2007; Elkins et al., 2013; Shaffer et al., 2013): If a doctor follows an AI when they have conflicting opinions, this may indicate that the doctor's skills are inferior to AI, causing him to be replaced in the future. The concern of being substituted in the future therefore makes the doctor reluctant to follow AI, creating the AI-aversion problem.

Although AI may be a substitute for doctors, it is also a complement. AI can make doctors more effective by providing valuable information about patients and increasing the precision of the doctor's diagnosis. The question then is how to better realize AI's potential and facilitate AI adoption by making it more of a complement rather than a substitute for doctors.

This paper addresses this question by focusing on one aspect of AI adoption: managing instances where the doctor and AI disagree. We observe that disagreements can arise for different reasons and, furthermore, the doctor may not know why they disagree. Whether AI can persuade the doctor then depends in part on how the doctor attributes the sources of disagreement. The AI can better persuade the doctor by inducing him to attribute the disagreement to the "right" source. One way to affect the attribution is through the design of the information available to the doctor. Our main result is that AI can be more persuasive when it reveals less information.

Specifically, we develop a framework that identifies two sources of disagreement in the context of disease diagnosis: *attention differences* and *comprehension differences*.[1] The attention difference arises when AI and the doctor observe different signs of the disease. For example, AI can identify an abnormality that the doctor overlooks. In this case, there is clear-cut evidence that the doctor made a mistake. The objectivity of the attention difference makes the doctor's diagnosis more accurate and complements the doctor. The attention difference, therefore, plays the role of a complement and facilitates AI persuasion.

The comprehension difference arises when AI and the doctor observe the same signs, but differ in how they assign importance to these signs. For example, AI may give a positive diagnosis by assigning a large weight to an abnormality. However, the doctor may think the abnormality is the result of a previous patient condition and assign a small weight to the abnormality. In this case, the difference of opinions is a matter of judgment. When the comprehension difference arises, the doctor may

[1]These two types of differences can also appear in many other contexts, such as bail decisions (Kleinberg et al., 2017), financial decisions (Kang and Kim, 2024), and purchasing decisions (Bundorf et al., 2019). In general, attention and comprehension differences correspond to the two key information constraints in decision-making (March, 1994, p. 9).

associate the change in opinion with the acknowledgment of a worse judgment. The subjectivity of the comprehension difference makes AI and the doctor compete to be the one with better judgment. The comprehension difference, therefore, plays the role of a substitute and makes AI persuasion less effective.

When the doctor disagrees with the AI's diagnosis, she may or may not know why they disagree. When the doctor can see which abnormalities are observed by AI, she knows for sure whether the disagreement is due to the attention difference or the comprehension difference. In this case, we call the AI *interpretable* because the doctor can interpret why they disagree. When the doctor cannot see which abnormalities are observed by AI, she no longer knows for sure the source of the disagreement. In this case, we refer to the AI as *uninterpretable*.

When AI is uninterpretable, the doctor can still assess why they disagree by performing *Bayesian attribution*. That is, she calculates the likelihood that the disagreement is due to the attention difference (and to the comprehension difference) according to the Bayes rule.

Our main result shows, perhaps paradoxically, that making AI uninterpretable can enhance persuasion. When AI is interpretable, doctors with good comprehension skills will not change their mind when they know that the diagnostic disagreement is due to the comprehension difference. When AI is uninterpretable, these doctors will always change their mind when AI disagrees with them.

To see why this is the case, note that when the AI is uninterpretable, the doctor *Bayesian attribute* the disagreement to attention differences with positive probability. Because the attention difference is more persuasive, a weighted average of comprehension and attention differences can persuade the doctor, even if the comprehension difference alone will not. Making AI uninterpretable, therefore, enhances persuasion by allowing the more persuasive attention difference to "subsidize" the less persuasive comprehension difference. In other words, the uninterpretable AI bundles the complement aspect of disagreements with the substitute aspect, and therefore makes the disagreement on average more of a complement than a substitute. We refer to this

mechanism as the *averaging effect*.

The weight the doctor assigns to the attention difference depends on her skill level. A doctor with lower attention skills (who is more likely to overlook abnormalities) is then more likely to attribute the disagreement to the attention difference. The higher the weight put on the attention difference, the more likely the doctor is to change her opinion. Consequently, making AI uninterpretable is particularly effective in improving persuasion when the doctor has lower attention skills. We refer to this mechanism as the *attribution effect*.

In addition to enhancing persuasion, making AI uninterpretable can also improve diagnostic accuracy if doctors have career concerns, that is, if they would like to be perceived as high-skilled. We demonstrate this point by considering a setting in which the doctors have the same attention skills, but they may have either high or low comprehension skills.

When the AI is interpretable, high-skilled doctors ignore AI because their comprehension skills are better. Low-skilled doctors, however, will ignore AI not because this improves diagnostic accuracy, but because they want to be perceived as the high-skilled ones and therefore would like to mimic the behaviors of high-skilled doctors. This leads to less accurate diagnoses.

Now, if the AI becomes uninterpretable and therefore more persuasive, high-skilled doctors will follow AI. This alleviates the career concerns of the low-skilled doctors, so they will also follow AI to make more accurate diagnoses. This result contrasts with the medical AI literature, which commonly views uninterpretability as a barrier to AI applications (e.g. He et al., 2019, and Kundu, 2021). Our result demonstrates that, by making the AI uninterpretable, we also enable the low-skilled doctors to change their minds without "losing face".

Related literature. Our paper is related to four broad literature. First, a growing literature investigates how AI complements workers, often examining how AI affects workers of different skills; see, for example, Gruber et al. (2020), Allen and Choudhury (2022), Brynjolfsson et al. (2023), Noy and Zhang (2023), Jia et al. (2024), and Wang

et al. (2024).[2] In these papers, the heterogeneity of workers is measured in a single dimension. Our paper highlights the multidimensional skills of doctors. We show that the effectiveness of AI assistance depends not just on the overall skills of the doctor, but also on the level of skills the doctor has on each dimension: the doctor benefits more when he has low attention skills than when he has low comprehension skills.

Second, our paper contributes to the literature on AI aversion. Tong et al. (2021) document that employees, once they know that the feedback is from an AI, will lower their trust in the feedback and more concern themselves with the risks of job replacement. Liu et al. (2023) find that drivers are less likely to follow AI's recommendations if they contradict past experiences and peers' actions. In the context of doctor-AI interaction, Jussupow et al. (2021) demonstrate that AI advice may undermine doctors' causal reasoning.[3] Lebovitz et al. (2022) report that doctors tend to explain disagreements away when AI recommendations conflict with their own judgments. Agarwal et al. (2023) document deviations from rational updating among doctors. This literature emphasizes the role of AI as a substitute for workers. We show that making AI uninterpretable allows the substitute role of AI to be bundled with its complement role, thereby facilitating AI adoption.

Third, our paper is related to the theoretical literature on human-AI interaction. This literature examines how human-AI interactions are affected by various underlying factors and design choices. Agrawal et al. (2018b, 2019) focus on different skills between AI and doctors. Athey et al. (2020) examine the optimal allocation of decision rights between agents and AI. Dai and Singh (2020) analyze how career concerns affect the acquisition of information by doctors. We add to this literature by studying how the interpretability of AI affects the way the doctor attributes the source of the disagreement.

Finally, this paper is related to the literature on information design; see Bergemann and Morris (2019) and Kamenica (2019) for general reviews on how the provision of information affects the behaviors of players. In the context of reputation concerns,

[2]See Ide and Talamas (2024) for the theoretical implications on the organization of knowledge work.
[3]Similar effects are also observed among financial analysts (Kang and Kim, 2024).

several studies have shown that non-transparency can improve welfare by disciplining reputational incentives (Prat, 2005; Levy, 2007; Ashworth and Shotts, 2010; Fox and Van Weelden, 2012; Fu and Li, 2014; De Moragas, 2022; Li, 2023). We extend this line of inquiry by examining the information design in the presence of explicit disagreements. Our study highlights that the benefits of limited information are larger when the doctors' have low attention skills.

2 Model

2.1 State Variables: Disease, Abnormality, and Critical Dimension

A patient consults a doctor for diagnosis of a potential disease. The disease manifests as an *abnormality* in one of two dimensions (e.g., medical tests). However, only one dimension is *critical*—an abnormality in this dimension indicates the presence of the disease. In contrast, an abnormality in the other dimension may or may not appear, regardless of whether the disease is present.[4]

Formally, we denote the disease status by $Z \in \{0,1\}$, where $Z = 1$ indicates its presence and $Z = 0$ its absence. Let L and R represent the two dimensions where an abnormality may appear, and W denote the critical dimension. Define $X = (X_L, X_R)$, where for $j \in \{L,R\}$, $X_j \in \{0,1\}$ represents the abnormality status in dimension j, with $X_j = 1$ indicating its presence and $X_j = 0$ its absence. Thus, X_W denotes the abnormality status in the critical dimension. We also use the compact notation X_{-W} to represent the abnormality status in the non-critical dimension. For example, if $W = L$, then $X_{-W} = X_R$, and if $W = R$, then $X_{-W} = X_L$.

The *state* of the world comprises the three variables (Z, X, W), whose relationship is illustrated in Figure 1. Their exact values are unknown and ex ante follow certain probability distributions.

Assumption 1. *(i) The patient has the disease with probability γ:* $\mathbf{P}(Z = 1) = \gamma$.

[4] According to Lebovitz et al. (2022), doctors ignore an unusual pattern for a patient if they have seen it long ago in prior imaging. For the patient, such abnormalities are deemed non-critical.

(ii) Either dimension is equally likely to be critical: $\mathbf{P}(W = L) = \mathbf{P}(W = R) = \frac{1}{2}$.

(iii) The abnormality status in the critical dimension is perfectly indicative of the disease status, whereas the abnormality status in the non-critical dimension is purely noise, independent of Z*:* $X_W = Z$ *and* $\lambda := \mathbf{P}(X_{-W} = 1) \in (0, 1)$.

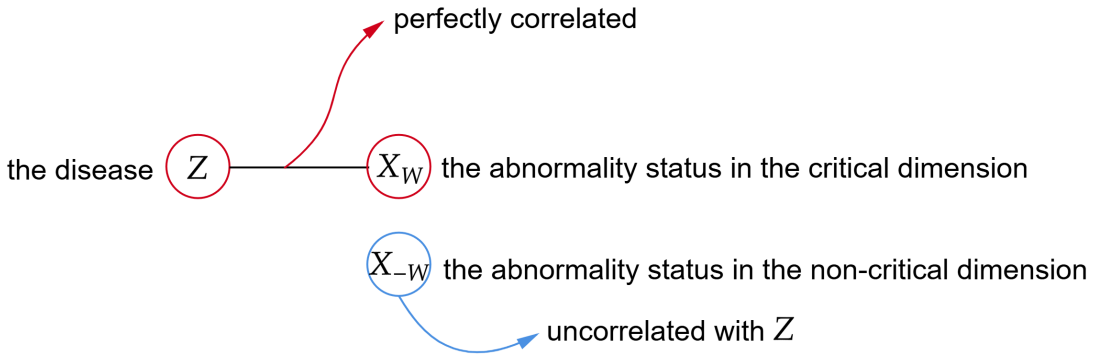


Figure 1: Disease status (Z), abnormality status (X), and critical dimension (W).

2.2 Signals for Diagnosis: Attention and Comprehension

The doctor relies on two types of signals for diagnosis of the disease: *attention* signal and *comprehension* signal. The former represents the doctor's observation of the abnormality status in both dimensions, whereas the latter reflects the doctor's judgment regarding which dimension is critical. Let $X^{\mathsf{Doc}} = (X_L^{\mathsf{Doc}}, X_R^{\mathsf{Doc}})$ denote the doctor's attention signal, where $X_j^{\mathsf{Doc}} \in \{0, 1\}$ for $j \in \{L, R\}$, and let $W^{\mathsf{Doc}} \in \{L, R\}$ denote her comprehension signal.

In addition to her own signals, the doctor has access to an AI providing diagnostic assistance. The AI also receives a pair of signals: an attention signal $X^{\mathsf{AI}} = (X_L^{\mathsf{AI}}, X_R^{\mathsf{AI}})$ and a comprehension signal W^{AI}. Our modeling of AI as an attention–comprehension pair is motivated by the prevalent use of artificial neural networks in AI technologies: X^{AI} represents a vector of input data, and W^{AI} reflects how AI weights different input dimensions.[5] Conditional on the state variables, the AI's signals are independent of the doctor's.

[5] A typical artificial neural network comprises multiple layers. The AI's attention signal in our model can be conceptualized as residing in the output layer, which is directly observable by the doctor. In practical medical applications, this is manifested as the abnormalities that the AI flags in medical images.

We call the AI *interpretable* if its signals are known to the doctor. In contrast, the AI is *uninterpretable* if the doctor cannot observe its signals, but only its diagnosis (which is derived from the AI's signals and will be elaborated later). We prefer the terms "interpretable" and "uninterpretable" over "transparent" and "nontransparent", because an AI can be uninterpretable even if it is partially transparent. For example, consider a radiologist diagnosing breast cancer by examining X-ray images. Even if a transparent AI flags several abnormalities, the radiologist may still find it uninterpretable as she does not know how the AI assigns importance to these abnormalities.

Moreover, we refer to the values of W^{Doc} and W^{AI} as the *doctor's critical dimension* and the *AI's critical dimension*, respectively. We make the following assumptions regarding the two types of signals for both the doctor and the AI:

Assumption 2. *(i) Neither the doctor nor the AI hallucinates:*[6] *for* $j \in \{L, R\}$, $\mathbf{P}(X_j^{\mathsf{Doc}} = 0 | X_j = 0) = 1$ *and* $\mathbf{P}(X_j^{\mathsf{AI}} = 0 | X_j = 0) = 1$.

(ii) Both the doctor and the AI may overlook an abnormality in either dimension: for $j \in \{L, R\}$, $\pi^{\mathsf{Doc}} := \mathbf{P}(X_j^{\mathsf{Doc}} = 1 | X_j = 1) \in [0, 1]$ *and* $\pi^{\mathsf{AI}} := \mathbf{P}(X_j^{\mathsf{AI}} = 1 | X_j = 1) \in [0, 1]$.

(iii) Both the doctor and the AI may miscomprehend the critical dimension: $p^{\mathsf{Doc}} := \mathbf{P}(W^{\mathsf{Doc}} = W | W) \in [\frac{1}{2}, 1]$ *and* $p^{\mathsf{AI}} := \mathbf{P}(W^{\mathsf{AI}} = W | W) \in [\frac{1}{2}, 1]$

2.3 Diagnostic Process

The doctor's diagnostic process with the AI's assistance is as follows:

1. Nature determines the disease status Z, abnormality status X, and critical dimension W.
2. The doctor receives signals $(X^{\mathsf{Doc}}, W^{\mathsf{Doc}})$ and then makes an *initial* diagnosis $D \in \{0, 1\}$.
3. The AI receives signals $(X^{\mathsf{AI}}, W^{\mathsf{AI}})$ and then makes a diagnosis $A \in \{0, 1\}$.
4. The doctor observes
   - $(X^{\mathsf{AI}}, W^{\mathsf{AI}})$ if the AI is interpretable, or

[6] A prominent challenge with large language models, the most recent development in AI technologies, is hallucination (Ji et al., 2023). Here, we assume no AI hallucination for simplicity. Our analysis can be generalized to cope with AI hallucination, allowing $\mathbf{P}(X_j^{\mathsf{AI}} = 0 | X_j = 0) \in [0, 1]$ (see Appendix C).

- A if the AI is uninterpretable.

The doctor then updates her belief about the disease status to make the *final* diagnosis $F \in \{0, 1\}$.

Each diagnosis (i.e., D, A, or F) is determined by the posterior probability of the presence of the disease given the corresponding information set. If this probability is greater than or equal to $\frac{1}{2}$, then the diagnosis is assigned a value of 1 and called *positive*. Otherwise, it is assigned a value of 0 and called *negative*.[7]

The quality of the doctor's initial diagnosis (i.e., without the AI's assistance) is fully determined by the accuracy of the pair of signals $(X^{\mathsf{Doc}}, W^{\mathsf{Doc}})$ relative to the underlying state variables (X, W). As a result, we refer to $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$ as the doctor's *attention skill* and *comprehension skill*, respectively. Likewise, $(\pi^{\mathsf{AI}}, p^{\mathsf{AI}})$ are termed the AI's attention skill and comprehension skill, respectively.

To simplify our exposition, we make the following assumptions regarding the values of γ, λ, $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$, and $(\pi^{\mathsf{AI}}, p^{\mathsf{AI}})$. (Recall that γ and λ are the ex ante the probability of the patient having the disease and the probability of an abnormality appearing in the non-critical dimension, respectively.)

Assumption 3. $\lambda \leq \gamma < \frac{1}{2}$.

Assumption 4. $(1/\gamma + 1/\lambda - 2)p^i + \pi^i > 1/\lambda$ *for* $i \in \{\mathsf{Doc}, \mathsf{AI}\}$.

Assumption 3 states that the disease occurs with probability $\gamma < \frac{1}{2}$. In addition, an abnormality is more likely to appear in the critical dimension, as it occurs there if and only if the patient has the disease under Assumption 1. Under Assumption 4, it can be shown that when the doctor (or the AI) only observes an abnormality in her non-critical dimension, the posterior probability of the disease status satisfies

$$\frac{\mathbf{P}(Z = 1 | X^i_{W^i} = 0, X^i_{-W^i} = 1)}{\mathbf{P}(Z = 0 | X^i_{W^i} = 0, X^i_{-W^i} = 1)} = \frac{\gamma[p^i\lambda(1 - \pi^i) + (1 - p^i)(1 - \lambda\pi^i)]}{\lambda(1 - \gamma)p^i} < 1,$$

[7]This process parallels classification tasks performed by AI technologies in fields such as computer vision and natural language processing. The threshold, however, is not necessarily $\frac{1}{2}$ but is typically adjusted between 0 and 1 for each specific task to balance the costs associated with false positive and false negative errors (Goodfellow et al., 2016). For simplicity, we assume the threshold is $\frac{1}{2}$ in our model.

which implies $\mathbf{P}(Z = 1|X^i_{W^i} = 0, X^i_{-W^i} = 1) < \frac{1}{2}$. Thus, in this case, neither the doctor nor the AI will make a positive diagnosis.

We maintain Assumptions 1–4 throughout the paper. Under these assumptions, the doctor will make a positive initial diagnosis if and only if the doctor observes an abnormality in her critical dimension. An analogous statement applies to the AI's diagnosis. Lemma 1 states this formally.

Lemma 1 (**Diagnosis Through Critical Dimension**). *The doctor makes a positive initial diagnosis (i.e., $D = 1$) if and only if $X^{\mathsf{Doc}}_{W^{\mathsf{Doc}}} = 1$. The AI makes a positive diagnosis (i.e., $A = 1$) if and only if $X^{\mathsf{AI}}_{W^{\mathsf{AI}}} = 1$.*

3 AI Persuasion

Will the doctor be persuaded to alter her initial diagnosis if it conflicts with the AI's diagnosis? How does the AI's interpretability affect its persuasiveness? This section explores these questions, beginning with the scenario where the doctor makes a negative initial diagnosis but the AI offers a positive diagnosis. We'll then briefly discuss the opposite case, which yields different results but similar insights.

3.1 AI Persuasion When $D = 0$ and $A = 1$

3.1.1 Interpretable AI

When the AI is interpretable, the doctor can observe the AI's attention–comprehension signals $(X^{\mathsf{AI}}, W^{\mathsf{AI}})$. This allows the doctor to pinpoint the source of disagreements, which may stem from either an *attention difference* or a *comprehension difference*.

Assuming the doctor has made a negative initial diagnosis, Lemma 1 indicates that she must have observed no abnormality in her critical dimension ($X^{\mathsf{Doc}}_{W^{\mathsf{Doc}}} = 0$). An attention difference arises when the AI observes an abnormality in the same dimension ($X^{\mathsf{AI}}_{W^{\mathsf{Doc}}} = 1$).

However, if the AI shares the same observation in the doctor's critical dimension

($X^{\mathsf{AI}}_{W^{\mathsf{Doc}}} = 0$), for it to make a positive diagnosis, it must have a different critical dimension ($W^{\mathsf{AI}} \neq W^{\mathsf{Doc}}$) and observe an abnormality there. This scenario represents a comprehension difference.[8]

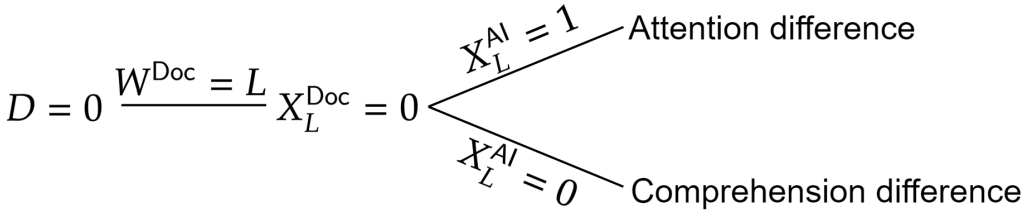


Figure 2: Two sources of disagreements when $D = 0$ and $A = 1$

Next, we summarize for both the attention difference and the comprehension difference how the AI can persuade the doctor based on the doctor's skills.

Lemma 2 (**Interpretable AI: Known Attribution**). *Suppose that the AI is interpretable and the diagnostic disagreement is given by $D = 0$ and $A = 1$. Then, the following hold:*

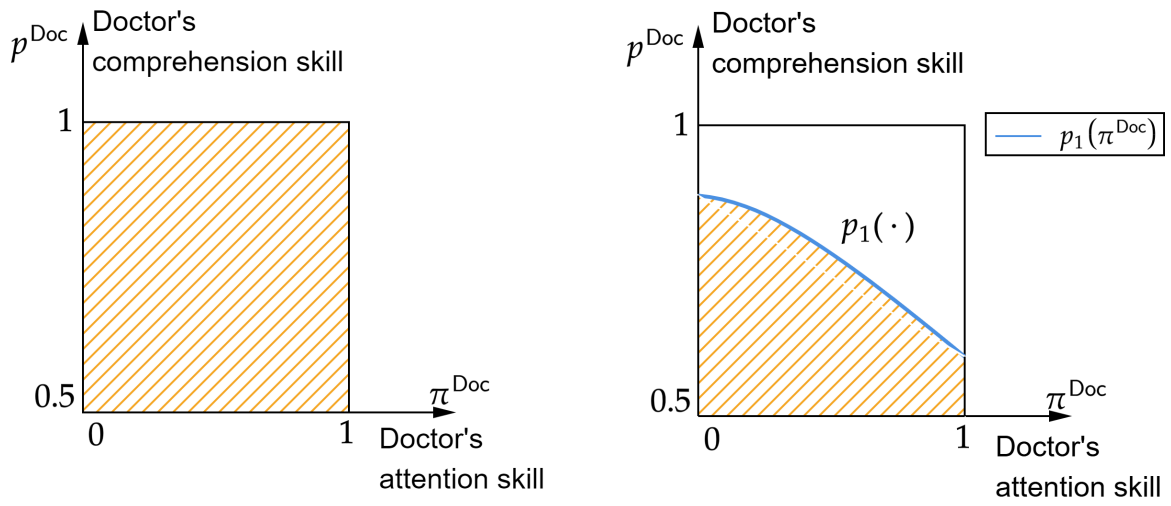
(i) When the attention difference occurs, the AI persuades the doctor to change her diagnosis from $D = 0$ to $F = 1$ regardless of her skill $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$.

(ii) When the comprehension difference occurs, the AI persuades the doctor if and only if her comprehension skill p^{Doc} is weakly below a threshold $p_1(\pi^{\mathsf{Doc}})$.

Part (i) of Lemma 2 shows that when an attention difference occurs, the AI persuades the doctor regardless of her skill level. This is because the attention difference provides clear evidence of the doctor's attentional mistake: she overlooked an abnormality in her critical dimension, which the AI has detected and corrected. By helping the doctor find the missed abnormality, attention difference serves as a complement to the doctor.

Part (ii) of Lemma 2 shows that when a comprehension difference occurs, the AI can persuade the doctor only if her comprehension skill p^{Doc} falls below the threshold $p_1(\pi^{\mathsf{Doc}})$. In this case, the main difference between the AI and the doctor lies in their judgment of the critical dimension. In contrast to the attention difference scenario, there is no clear-cut evidence of a mistake by the doctor: the differences are in opinions

[8] Alternatively, one could define the attention difference as a discrepancy in attention signals (i.e., $X^{\mathsf{Doc}} \neq X^{\mathsf{AI}}$), and the comprehension difference as a discrepancy in comprehension signals (i.e., $W^{\mathsf{Doc}} \neq W^{\mathsf{AI}}$). However, we avoid this approach as it would permit both differences to occur simultaneously, thereby complicating the analysis of AI persuasiveness.



(a) Attention persuasion (b) Comprehension persuasion

Figure 3: The range of doctors persuaded by the interpretable AI

rather than in facts. The comprehension difference, therefore, makes AI a substitute for AI, obstructing the effectiveness of persuasion. When the comprehension difference occurs, the doctor is persuaded only if her comprehension skill is sufficiently low. In particular, the threshold $p_1(\pi^{\text{Doc}})$ is downward sloping because a doctor with better attention skills is less likely to overlook abnormalities and thus more resistant to changing her initial diagnosis.

A key insight from our analysis of interpretable AI is that attention differences are more persuasive than comprehension differences: the attention difference makes the AI a complement, and the comprehension difference makes it a substitute. This finding helps explain why doctors may resist AI recommendations even after receiving explanations (Clement et al., 2021). It also sheds light on the observation by Lebovitz et al. (2022) that radiologists tend to follow AI for diagnosing lung cancer but disregard it for diagnosing breast cancer. A lung cancer diagnosis is primarily based on detecting nodules, and once identified, interpretation is relatively straightforward. This makes spotting abnormalities crucial, while differences in comprehension are less significant. AI thus complements doctors by providing valuable attention signals without introducing substantial comprehension differences. In contrast, breast cancer presents a wider variety of abnormalities, and even when detected, their interpretation can vary. This complexity elevates the importance of comprehension in diagnosis. Conse-

quently, AI's role in creating comprehension differences makes it more of a substitute, diminishing the persuasiveness of AI.

As AI can be both a substitute and a complement, the question is how to find ways to make it better complement the doctor. We show next that this can be achieved by making AI uninterpretable.

3.1.2 Uninterpretable AI

When AI is uninterpretable, the doctor observes only the AI's diagnosis, not the abnormalities it spots. The doctor's information set is given by $\mathcal{I}_{0,1}^{\mathsf{Doc}} = (X^{\mathsf{Doc}}, W^{\mathsf{Doc}}, D = 0, A = 1)$. When the doctor disagrees with AI, she carries out *Bayesian attribution*. That is, she applies Bayes' rule to attribute the disagreement to attention and comprehension differences. Define the conditional probability of the attention difference as

$$\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) := \mathbf{P}(X_{W^{\mathsf{Doc}}}^{\mathsf{AI}} = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}})$$

and the conditional probability of the comprehension difference as

$$\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) := \mathbf{P}(X_{W^{\mathsf{Doc}}}^{\mathsf{AI}} = 0|\mathcal{I}_{0,1}^{\mathsf{Doc}}).$$

The doctor can then decompose the conditional probability of the disease's presence as a weighted average of the two sources of disagreements:

$$\begin{aligned}\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}}) = \mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \cdot \mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) \\ + \mathbf{P}(\mathsf{Comp}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \cdot \mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\mathsf{Doc}}).\end{aligned} \tag{1}$$

She is persuaded to change her initial diagnosis if and only if $\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$. The following corollary and its implications provide foundations for our Bayesian attribution decomposition (Equation (1)). These results will guide our characterization of how the uninterpretable AI persuades the doctor.

Corollary 1 (Attention Difference Is More Persuasive). *The AI is more persuasive when*

the diagnostic disagreement stems from an attention difference rather than a comprehension difference: $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) > \frac{1}{2}$ *for any* $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$*, but* $\mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$ *only if* $p^{\mathsf{Doc}} \leq p_1(\pi^{\mathsf{Doc}})$.

To understand this corollary, consider a scenario where the doctor, facing an uninterpretable AI, contemplates what belief she would hold and what role the AI would be playing if she were aware of the exact source of the diagnostic disagreement. In such a situation, her belief is formed as though the AI were interpretable. Thus, Corollary 1 directly follows from Lemma 2. Because the attention difference is persuasive regardless of the doctor's skills (Part (i) of Lemma 2), $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) > \frac{1}{2}$ for any $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$. This inequality is strict because integrating the AI's attention signal with the doctor's may reveal the disease by showing $X_L = X_R = 1$. However, the comprehension difference is persuasive only if the doctor's comprehension skill is sufficiently low (Part (ii) of Lemma 2). Hence, $\mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$ only if $p^{\mathsf{Doc}} \leq p_1(\pi^{\mathsf{Doc}})$.

When AI is uninterpretable, the exact source of disagreement remains unknown to the doctor. An implication of Corollary 1 is that the role of an uninterpretable AI — whether it is a complement or a substitute — depends on the weight the doctor assigns to the attention difference in her Bayesian attribution, i.e., $\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}})$. As long as the attention difference receives sufficient attribution, the uninterpretable AI will resemble a complement to the doctor. The following proposition formalizes this implication.

Proposition 1 (Uninterpretability Enhances Persuasion). *Suppose that the AI is uninterpretable and the diagnostic disagreement is given by* $D = 0$ *and* $A = 1$*. Then, the following hold:*

(i) **Threshold for persuasion:** *There exists a threshold,* $p_2(\pi^{\mathsf{Doc}})$*, such that the AI persuades the doctor whenever* $p^{\mathsf{Doc}} \leq p_2(\pi^{\mathsf{Doc}})$.

(ii) **Averaging effect:** $p_2(\pi^{\mathsf{Doc}}) \geq p_1(\pi^{\mathsf{Doc}})$*, and the inequality is strict if* $\pi^{\mathsf{Doc}} < 1$.

(iii) **Attribution effect:** *As* π^{Doc} *decreases,* $p_2(\pi^{\mathsf{Doc}}) - p_1(\pi^{\mathsf{Doc}})$ *increases.*

Part (i) of Proposition 1 shows that there exists a threshold, $p_2(\pi^{\mathsf{Doc}})$, such that if the doctor's comprehension skill is at or below this threshold, the uninterpretable AI

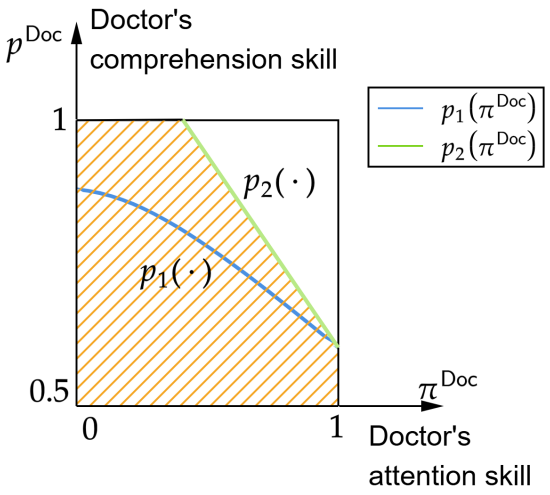


Figure 4: The range of doctors persuaded by the uninterpretable AI

successfully persuades the doctor to make a positive final diagnosis. In this case, the AI complements the doctor.

In particular, Part (ii) of Proposition 1 shows that $p_2(\pi^{\text{Doc}}) \geq p_1(\pi^{\text{Doc}})$ (Figure 4), indicating that when her comprehension skill falls between $p_1(\pi^{\text{Doc}})$ and $p_2(\pi^{\text{Doc}})$, the doctor remains unpersuaded by the interpretable AI but is persuaded by the uninterpretable AI. This phenomenon arises from what we term the *averaging effect*. With an uninterpretable AI, the doctor averages the persuasiveness of attention and comprehension differences. Since an attention difference is inherently more persuasive than a comprehension difference, this averaging effect enhances the doctor's inclination to follow the AI's recommendation. In other words, making AI less interpretable turns AI into a complement for some doctors.

Equation (1) formally captures the averaging effect, decomposing the conditional probability of the disease into components based on attention and comprehension differences. By Corollary 1, when p^{Doc} slightly exceeds $p_1(\pi^{\text{Doc}})$, we have $\mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\text{Doc}}) < \frac{1}{2}$, indicating that the comprehension difference alone is not persuasive. However, since $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\text{Doc}}) > \frac{1}{2}$, the averaging effect can still result in $\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\text{Doc}}) \geq \frac{1}{2}$. Therefore, even when $p^{\text{Doc}} > p_1(\pi^{\text{Doc}})$, the uninterpretable AI retains the potential to persuade the doctor.

Further, Part (iii) of Proposition 1 shows that a decrease in the doctor's attention skill can widen the gap between $p_2(\pi^{\text{Doc}})$ and $p_1(\pi^{\text{Doc}})$ by increasing the attribution to

attention differences. This result arises because, in Equation (1), the weights the doctor assigns to attention and comprehension differences depend on her attention skill. We call this mechanism the *attribution effect*.

Consider the scenario where the doctor observes an abnormality in her non-critical dimension, i.e., $X^{\text{Doc}} = (0,1)$ or $X^{\text{Doc}} = (1,0)$. In this scenario, a lower attention skill implies a higher likelihood of the doctor overlooking abnormalities in her critical dimension. Consequently, as her attention skill decreases, she attributes the diagnostic disagreement more to the attention difference, becoming more susceptible to persuasion. Due to this attribution effect, making AI uninterpretable proves more effective in enhancing persuasion when the doctor is less attentive.

3.2 AI Persuasion When $D = 1$ and $A = 0$

Beyond the diagnostic disagreement examined in Section 3.1, the doctor may encounter a scenario where her initial diagnosis is positive ($D = 1$) and the AI's diagnosis is negative ($A = 0$). In this case, the attention difference (i.e., $W^{\text{AI}} = W^{\text{Doc}}$) is completely unpersuasive, while the comprehension difference (i.e., $W^{\text{AI}} \neq W^{\text{Doc}}$) is persuasive if and only if the doctor observes only one abnormality ($X^{\text{Doc}} \neq (1,1)$), and her comprehension skill p^{Doc} falls below a threshold $p_3(\pi^{\text{Doc}})$. Given the lack of persuasiveness in the attention difference, the interpretable AI may fail to convince the doctor.

However, with an uninterpretable AI, there exists another threshold, $p_4(\pi^{\text{Doc}})$, such that if $p^{\text{Doc}} < p_4(\pi^{\text{Doc}})$, the AI always persuades the doctor provided $X^{\text{Doc}} \neq (1,1)$. This results from the same averaging effect discussed previously. Interestingly, as the comprehension difference is now more persuasive than the attention difference, the attribution effect is reversed, enhancing the persuasion of more attentive doctors.

Appendix A offers a detailed analysis of this case. Notably, for any π^{Doc}, $p_4(\pi^{\text{Doc}}) \leq p_3(\pi^{\text{Doc}}) \leq p_1(\pi^{\text{Doc}}) \leq p_2(\pi^{\text{Doc}})$. This ranking allows us to treat the two kinds of diagnostic disagreements in Sections 3.1 and 3.2 as independent cases and examine them separately.

4 Career Concerns

Thus far, we have assumed that doctors maximize diagnostic accuracy. In practice, they also care about their reputation (Chan et al., 2022). We now examine the doctor's career-concern problem, where doctors are concerned about their reputation. We focus on the reputation for strong comprehension skills because interpreting abnormalities requires more expertise than identifying them.

We follow Li (2023) to model the doctor's career concern. Consider two types of doctors who differ in their comprehension skills. A high-type doctor possesses perfect comprehension ($p^{\text{Doc}_H} = 1$), while a low-type has imperfect comprehension ($p^{\text{Doc}_L} < p^{\text{AI}} \leq 1$). Ex ante, the probability of a doctor being high-type is τ, and the probability of her being low-type is $1 - \tau$. Only the doctor herself is aware of her type.

The doctor's objective is to maximize an evaluator's (e.g., a manager's) belief about her type, rather than her diagnostic accuracy.[9] The evaluator observes the doctor's signals,[10] her initial and final diagnoses, and AI information. Formally, when the AI is interpretable, the evaluator observes $(X^{\text{Doc}}, W^{\text{Doc}}, D, X^{\text{AI}}, W^{\text{AI}}, A, F)$, and when the AI is uninterpretable, the evaluator observes $(X^{\text{Doc}}, W^{\text{Doc}}, D, A, F)$. Following Li (2023), we focus on the equilibrium where high-type doctors make efficient diagnoses.

Due to career concerns, low-type doctors may make an inefficient diagnosis. To illustrate this, consider a scenario where $X^{\text{Doc}} = X^{\text{AI}} = (0, 1)$, $W^{\text{Doc}} = L$, $W^{\text{AI}} = R$, and the AI is interpretable. In this scenario, both the doctor and AI observe an abnormality in dimension R, but only the AI perceives it as critical. Because the low-type doctor's comprehension skill is lower than the AI's, and the high-type doctor's comprehension skill is perfect, efficiency would dictate that the low-type doctor change her comprehension, while the high-type doctor maintain hers. However, in an attempt to emulate the high type, the low-type doctor may insist on her initial comprehension.

[9] This assumption represents the extreme case where doctors have no concerns for diagnostic accuracy. In general, we can consider cases when doctors care about both accuracy and reputation. Our result would then be that uninterpretability can improve the doctor's diagnostic accuracy as long as her career concern is sufficiently strong.

[10] We consider these signals verifiable because, in practice, the doctor needs to record the signs she observes and state her comprehension to patients (or colleagues) at the initial diagnostic stage.

This strategic behavior reduces diagnostic accuracy.

Recall from Proposition 1 that making the AI uninterpretable can motivate the doctor to follow the AI's recommendation by transforming the AI into a complement. Building on this result, we show that AI's uninterpretability can effectively manage the doctor's career concerns.

Proposition 2 (Uninterpretability Improves Accuracy). *Suppose* $p^{\mathsf{Doc}_H} < p_2(\pi^{\mathsf{Doc}})$ *and* $p^{\mathsf{Doc}_L} > p_3(\pi^{\mathsf{Doc}})$. *Then, there exists* $\overline{\tau} \in (0,1)$ *such that if* $\tau < \overline{\tau}$, *the accuracy of the final diagnosis is higher when the AI is uninterpretable compared to when it is interpretable.*

This proposition shows that uninterpretability can enhance diagnostic accuracy when the doctor has career concerns. To illustrate, consider the previous example where $X^{\mathsf{Doc}} = X^{\mathsf{AI}} = (0,1)$, $W^{\mathsf{Doc}} = L$, and $W^{\mathsf{AI}} = R$. With an interpretable AI, both high-type and low-type doctors adhere to their initial comprehension and diagnosis. In contrast, when the AI is uninterpretable, high-type doctors are persuaded due to the increased persuasiveness of uninterpretable AI (as shown in Proposition 1). Then, to emulate the high types, low-type doctors are also persuaded. This change improves the diagnostic accuracy of low-type doctors.[11] When the ex-ante probability of the doctor being a low type is sufficiently high ($\tau < \overline{\tau}$), the change also improves the average diagnostic accuracy across all doctors.

Our result dovetails with recent empirical evidence on AI aversion. DeStefano et al. (2022) demonstrate that reducing AI interpretability can enhance performance by encouraging decision-makers to accept AI recommendations. They also suggest that decision-makers are more likely to accept AI recommendations when respected peers are involved in the AI development and testing process. Similarly, Kawaguchi (2021) shows that low-performing workers are more willing to follow AI recommendations when their own forecasts are incorporated into the AI system. These behavioral responses can be attributed to obfuscated attribution: as AI integrates human information, it becomes less clear whether decision-makers are deferring to AI or the

[11] Due to AI's uninterpretability, doctors incur a loss in combining the AI's attention signal. However, when $p^{\mathsf{Doc}} < p^{\mathsf{AI}}$, this loss is strictly dominated by the benefit mentioned in the text (see Lemma B.1 and the proof of Proposition 2 in Appendix B).

integrated human input when following AI recommendations. This intuition has important implications for human-AI collaboration. By making AI less interpretable, it can help preserve human dignity in decision-making processes, facilitating more efficient collaboration between humans and AI systems.

5 Conclusion

This paper develops a framework to study how AI persuades doctors when they have different diagnoses of diseases. We highlight that disagreements between doctors and AI can arise from two sources. The attention difference is objective and plays the role of a complement, helping the AI to persuade. In contrast, the comprehension difference is subjective and plays the role of a substitute, making AI less persuasive.

We show that AI's interpretability affects how the doctor attributes these sources and, therefore, the doctor's willingness to change her mind. An uninterpretable AI can become more persuasive by allowing the doctor to attribute the disagreement partly to the attention difference. This effect is stronger when doctors have low skills in detecting abnormalities. We also show that making AI uninterpretable can increase diagnostic accuracy when doctors have career concerns.

We have kept the model's complexity to a minimum by abstracting away various features commonly seen in practice. For instance, we keep the information structure simple by assuming that AI does not produce hallucinations. Furthermore, we have not considered the motivational issues of the doctor. We show in Appendix C that an uninterpretable AI can still be more persuasive, even if AI may hallucinate in observing abnormalities. Appendix D allows the doctor to draw a costly additional signal about the disease when she disagrees with the AI. We show that making the AI uninterpretable has the extra benefit of motivating the doctor to acquire information.

Our paper explores disagreements between AI and doctors, but the implications of Bayesian attribution extend further. For instance, when a subordinate disagrees with a leader, it can be unclear if the disagreement reflects genuine opinion or an

attempt to undermine authority. If the leader perceives it as the latter, the subordinate may withhold valuable information, harming collaboration. Organizations should design protocols that promote more positive attribution from leaders to encourage open communication.

Finally, we discuss how making AI uninterpretable allows us to combine its substitute and complementary aspects, positioning AI as a complement to doctors. Other professions, like accounting, consulting, and law, face similar challenges: while AI can enhance effectiveness, it also poses a risk of replacement. This raises the question of how to design jobs by restructuring and bundling tasks to ensure that AI is a complement rather than a substitute. Successful job redesign promotes greater adoption of AI technologies. This is a vital area that warrants further research.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology." Working Paper 31422, National Bureau of Economic Research.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018a. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018b. *Prediction, Judgment, and Complexity: A Theory of Decision-Making and Artificial Intelligence*. 89–110, University of Chicago Press.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2022. *Power and Prediction: The Disruptive economics of Artificial Intelligence*. Harvard Business Press.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Exploring the Impact of Artificial Intelligence: Prediction versus judgment." *Information Economics and Policy* 47 1–6.

Allen, Ryan, and Prithwiraj (Raj) Choudhury. 2022. "Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion." *Organization Science* 33 (1): 149–169.

Arkes, Hal R., Victoria A. Shaffer, and Mitchell A. Medow. 2007. "Patients Derogate Physicians Who Use a Computer-Assisted Diagnostic Aid." *Medical Decision Making* 27 (2): 189–202.

Ashworth, Scott, and Kenneth W. Shotts. 2010. "Does Informative Media Commentary Reduce Politicians' Incentives to Pander?" *Journal of Public Economics* 94 (11): 838–847.

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans. 2020. "The Allocation of Decision Authority to Human and Artificial Intelligence." *AEA Papers and Proceedings* 110 80–84.

Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20 1–12, New York, NY, USA: Association for Computing Machinery.

Bergemann, Dirk, and Stephen Morris. 2019. "Information Design: A Unified Perspective." *Journal of Economic Literature* 57 (1): 44–95.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2023. "Generative AI at Work." Working Paper 31161, National Bureau of Economic Research.

Brynjolfsson, Erik, Chong Wang, and Xiaoquan Zhang. 2021. "The Economics of IT and Digitization: Eight Questions for Research." *MIS Quarterly* 45 (1): 473–477.

Bundorf, M. Kate, Maria Polyakova, and Ming Tai-Seale. 2019. "How do Humans Interact with Algorithms? Experimental Evidence from Health Insurance." Working Paper 25976, National Bureau of Economic Research.

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33 (2): 220–239.

Cao, Sean, Wei Jiang, Junbo Wang, and Baozhong Yang. 2024. "From Man vs. Machine to Man + Machine: The art and AI of stock analyses." *Journal of Financial Economics* 160 103910.

Chan, David C, Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *The Quarterly Journal of Economics* 137 (2): 729–783.

Chen, Zenan, and Jason Chan. 2023. "Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise." *Available at SSRN 4575598*.

Clement, Jeffrey, Yuqing Ren, and Shawn Curley. 2021. "Increasing System Transparency About Medical AI Recommendations May Not Improve Clinical Experts' Decision Quality." *Available at SSRN 3961156*.

Dai, Tinglong, and Shubhranshu Singh. 2020. "Conspicuous by Its Absence: Diagnostic Expert Testing Under Uncertainty." *Marketing Science* 39 (3): 540–563.

De Freitas, Julian, Stuti Agarwal, Bernd Schmitt, and Nick Haslam. 2023. "Psychological factors underlying attitudes toward AI tools." *Nature Human Behaviour* 7 (11): 1845–1854.

De Moragas, Antoni-Italo. 2022. "Disclosing decision makers' private interests." *European Economic Review* 150 104282.

Dell'Acqua, Fabrizio, Bruce Kogut, and Patryk Perkowski. 2023a. "Super Mario Meets AI: Experimental Effects of Automation and Skills on Team Performance and Coordination." *Review of Economics and Statistics,* forthcoming.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R Mollick et al. 2023b. "Navigating the Jagged Technological Frontier: Field Rxperimental Evidence of The Effects of AI

on Knowledge Worker Productivity and Quality." *Harvard Business School Technology & Operations Management Unit Working Paper* (24-013): .

DeStefano, Timothy, Katherine Kellogg, Michael Menietti, and Luca Vendraminelli. 2022. "Why Providing Humans with Interpretable Algorithms May, Counterintuitively, Lead to Lower Decision-making Performance." *MIT Sloan Research Paper*.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology. General* 144 (1): 114–126.

Dranove, David, and Craig Garthwaite. 2022. "Artificial Intelligence, the Evolution of the Healthcare Value Chain, and the Future of the Physician." Working Paper 30607, National Bureau of Economic Research.

Elkins, Aaron C., Norah E. Dunbar, Bradley Adame, and Jay F. Nunamaker. 2013. "Are Users Threatened by Credibility Assessment Systems?" *Journal of Management Information Systems* 29 (4): 249–261.

Fox, Justin, and Richard Van Weelden. 2012. "Costly Transparency." *Journal of Public Economics* 96 (1): 142–150.

Fu, Qiang, and Ming Li. 2014. "Reputation-concerned Policy Makers and Institutional Status quo Bias." *Journal of Public Economics* 110 15–25.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Grossmann, Igor, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. "AI and the Transformation of Social Science Research." *Science* 380 (6650): 1108–1109.

Gruber, Jonathan, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad. 2020. "Managing Intelligence: Skilled Experts and AI in Markets for Complex Products." Working Paper 27038, National Bureau of Economic Research.

He, Jianxing, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. "The Practical Implementation of Artificial Intelligence Technologies in Medicine." *Nature Medicine* 25 (1): 30–36.

Ide, Enrique, and Eduard Talamas. 2024. "Artificial Intelligence in the Knowledge Economy." *Available at arXiv:2312.05481*.

Ji, Ziwei, Nayeon Lee, Rita Frieske et al. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55 (12): Article 248.

Jia, Nan, Xueming Luo, Zheng Fang, and Chengcheng Liao. 2024. "When and How Artificial Intelligence Augments Employee Creativity." *Academy of Management Journal* 67 (1): 5–32.

Jussupow, Ekaterina, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. 2021. "Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence." *Information Systems Research* 32 (3): 713–735.

Kamenica, Emir. 2019. "Bayesian Persuasion and Information Design." *Annual Review of Economics* 11 249–272.

Kang, Xi, and Hyunjin Kim. 2024. "Predictive Algorithms and Decision-making: Evidence from a Field Experiment." working paper, Vanderbilt University.

Kawaguchi, Kohei. 2021. "When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business." *Management Science* 67 (3): 1670–1695.

Kim, Hyunjin, Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2024. "Decision authority and the returns to algorithms." *Strategic Management Journal* 45 (4): 619–648.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–293.

Kundu, Shinjini. 2021. "AI in Medicine Must Be Explainable." *Nature Medicine* 27 (8): 1328–1328.

Lebovitz, Sarah, Hila Lifshitz-Assaf, and Natalia Levina. 2022. "To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis." *Organization Science* 33 (1): 126–148.

Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review* 97 (1): 150–168.

Li, Hanzhe. 2023. "Transparency and Policymaking with Endogenous Information Provision." *Available at arXiv:2204.08876*.

Liu, Meng, Xiaocheng Tang, Siyuan Xia, Shuo Zhang, Yuting Zhu, and Qianying Meng. 2023. "Algorithm Aversion: Evidence from Ridesharing Drivers." *Management Science,* forthcoming.

Longoni, Chiara, Andrea Bonezzi, and Carey K Morewedge. 2019. "Resistance to Medical Artificial Intelligence." *Journal of Consumer Research* 46 (4): 629–650.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu. 2019. "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases." *Marketing Science* 38 (6): 937–947.

Mahmud, Hasan, A.K.M. Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. "What Influences Algorithmic Decision-making? A Systematic Literature Review on Algorithm Aversion." *Technological Forecasting and Social Change* 175 121390.

March, James G. 1994. *A Primer on Decision Making: How Decisions Happen*. Simon and Schuster.

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics* 137 (2): 679–727.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Science* 381 (6654): 187–192.

Otis, Nicholas, Rowan P Clarke, Solene Delecourt, David Holtz, and Rembrand Koning. 2023. "The Uneven Impact of Generative AI on Entrepreneurial Performance." *Available at SSRN 4671369*.

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95 (3): 862–877.

Shaffer, Victoria A., C. Adam Probst, Edgar C. Merkle, Hal R. Arkes, and Mitchell A. Medow. 2013. "Why Do Patients Derogate Physicians Who Use a Computer-Based Diagnostic Support System?" *Medical Decision Making* 33 (1): 108–118.

Tong, Siliang, Nan Jia, Xueming Luo, and Zheng Fang. 2021. "The Janus Face of Artificial Intelligence Feedback: Deployment versus Disclosure Effects on Employee Performance." *Strategic Management Journal* 42 (9): 1600–1631.

Vanneste, Bart S., and Phanish Puranam. 2024. "Artificial Intelligence, Trust, and Perceptions of Agency." *Academy of Management Review,* forthcoming.

Wang, Weiguang, Guodong (Gordon) Gao, and Ritu Agarwal. 2024. "Friend or Foe? Teaming Between Artificial Intelligence and Workers with Variation in Experience." *Management Science* 70 (9): 5753–5775.

Yin, Yidan, Nan Jia, and Cheryl J. Wakslak. 2024. "AI Can Help People Feel heard, But An AI Label Diminishes This Impact." *Proceedings of the National Academy of Sciences* 121 (14): e2319112121.

Yu, Feiyang, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. 2024. "Heterogeneity and Predictors of The Effects of AI Assistance on Radiologists." *Nature Medicine* 30 (3): 837–849.

A More on AI Persuasion When $D = 1$ and $A = 0$

We provide additional discussion on the scenario where the doctor's initial diagnosis is positive ($D = 1$) and the AI's diagnosis is negative ($A = 0$). In this case, the doctor observes an abnormality in her critical dimension, but the AI observes no abnormality in its critical dimension. The disagreement, once again, stems from two sources: (1) $W^{\mathsf{AI}} = W^{\mathsf{Doc}}$, referred to as *attention difference*, and (2) $W^{\mathsf{AI}} \neq W^{\mathsf{Doc}}$, referred to as *comprehension difference*. The following lemma connects these two sources to AI persuasion.

Lemma A.1. *Suppose that the AI is interpretable and the diagnostic disagreement is given by $D = 1$ and $A = 0$. Then, the following hold:*

(i) When the attention difference occurs, the AI cannot persuade the doctor to change her diagnosis from $D = 1$ to $F = 0$.

(ii) When the comprehension difference occurs, the AI persuades the doctor if and only if $X^{\mathsf{Doc}} \neq (1, 1)$, and her comprehension skill, p^{Doc}, is weakly below a threshold $p_3(\pi^{\mathsf{Doc}})$.

Lemma A.1 provides conditions under which the interpretable AI can persuade the doctor. When the doctor makes a positive initial diagnosis, she observes an abnormality in her critical dimension. To change her diagnosis, the AI must switch her comprehension to a dimension where no abnormality is observed. Therefore, if the doctor is persuaded, the AI must perceive a different critical dimension from the doctor ($W^{\mathsf{AI}} \neq W^{\mathsf{Doc}}$) in which the doctor observes no abnormality ($X^{\mathsf{Doc}} \neq (1, 1)$). Furthermore, the doctor's comprehension skill must be sufficiently low ($p^{\mathsf{Doc}} < p_3(\pi^{\mathsf{Doc}})$) so that she defers to the AI's comprehension.

Next, suppose that the AI is uninterpretable. In this case, the doctor's information set is given by $\mathcal{I}_{1,0}^{\mathsf{Doc}} = (X^{\mathsf{Doc}}, W^{\mathsf{Doc}}, D = 1, A = 0)$. The doctor can decompose her belief similarly to Equation (1), considering the attribution of the disagreement and the conditional probability of the disease given the information set $\mathcal{I}_{1,0}^{\mathsf{Doc}}$. The following corollary states that the comprehension difference is more persuasive than the attention difference.

Corollary A.1. *The AI is more persuasive with the comprehension difference than the attention difference:* $\mathbf{P}(Z = 0|\mathsf{Comp}, \mathcal{I}_{1,0}^{\mathsf{Doc}}) \geq \frac{1}{2}$ *if* $X^{\mathsf{Doc}} \neq (1,1)$ *and* $p^{\mathsf{Doc}} \leq p_3(\pi^{\mathsf{Doc}})$*, but* $\mathbf{P}(Z = 0|\mathsf{Atten}, \mathcal{I}_{1,0}^{\mathsf{Doc}}) < \frac{1}{2}$.

Finally, we show how AI's uninterpretability enhances persuasion through the averaging effect and the attribution effect.

Proposition A.1. *Suppose that the AI is uninterpretable and the diagnostic disagreement is given by* $D = 1$ *and* $A = 0$*. Then, if* $X^{\mathsf{Doc}} = (1,1)$*, the AI cannot persuade the doctor. If* $X^{\mathsf{Doc}} \neq (1,1)$*, the following hold:*

(i) **Threshold for persuasion:** *There exists a threshold,* $p_4(\pi^{\mathsf{Doc}})$*, such that if* $p^{\mathsf{Doc}} \leq p_4(\pi^{\mathsf{Doc}})$*, the AI persuades the doctor. In particular,* $p_4(\pi^{\mathsf{Doc}}) \leq p_3(\pi^{\mathsf{Doc}})$.

(ii) **Averaging effect:** *When* $p^{\mathsf{Doc}} \leq p_4(\pi^{\mathsf{Doc}})$*, the AI can persuade the doctor, but it would not if the AI were interpretable.*

(iii) **Attribution effect:** *As* π^{Doc} *increases,* $p_3(\pi^{\mathsf{Doc}}) - p_4(\pi^{\mathsf{Doc}})$ *decreases.*

Similar to Proposition 1, where the diagnostic disagreement is given by $D = 0$ and $A = 1$, Proposition A.1 also characterizes persuasion through the two effects of AI's uninterpretability. However, the attribution effect is reversed. This reversal arises because the attention difference is actually less persuasive than the comprehension difference when the diagnostic disagreement is given by $D = 1$ and $A = 0$. Nevertheless, as mentioned in the text, AI's uninterpretability still enhances persuasion by averaging the two sources of disagreements.

B Proofs

Proof of Lemma 1. Without loss of generality, suppose $W^{\mathsf{Doc}} = L$. There are four cases to consider about the doctor's attention signal: (i) $X^{\mathsf{Doc}} = (1,1)$, (ii) $X^{\mathsf{Doc}} = (0,0)$, (iii) $X^{\mathsf{Doc}} = (1,0)$, and (iv) $X^{\mathsf{Doc}} = (0,1)$.

In (i), $\mathbf{P}(Z = 1|X^{\mathsf{Doc}}, W^{\mathsf{Doc}}) = 1$. Thus $D = 1$.

In (ii), $\dfrac{\mathbf{P}(Z = 1|X^{\mathsf{Doc}}, W^{\mathsf{Doc}})}{\mathbf{P}(Z = 0|X^{\mathsf{Doc}}, W^{\mathsf{Doc}})} = \dfrac{\gamma(1 - \pi^{\mathsf{Doc}})}{1 - \gamma} < 1$, where the inequality follows from $\gamma < \frac{1}{2}$. Thus $D = 0$.

In (iii), $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{Doc}},W^{\mathsf{Doc}})}{\mathbf{P}(Z=0|X^{\mathsf{Doc}},W^{\mathsf{Doc}})}=\dfrac{\gamma[p^{\mathsf{Doc}}(1-\lambda\pi^{\mathsf{Doc}})+(1-p^{\mathsf{Doc}})\lambda(1-\pi^{\mathsf{Doc}})]}{(1-\gamma)(1-p^{\mathsf{Doc}})\lambda}\geq 1$, where the inequality follows from $\gamma(1-\lambda\pi^{\mathsf{Doc}})\geq(1-\gamma)\lambda$ and $p^{\mathsf{Doc}}\geq 1-p^{\mathsf{Doc}}$. Thus $D=1$.

In (iv), $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{Doc}},W^{\mathsf{Doc}})}{\mathbf{P}(Z=0|X^{\mathsf{Doc}},W^{\mathsf{Doc}})}=\dfrac{\gamma[p^{\mathsf{Doc}}(1-\pi^{\mathsf{Doc}})\lambda+(1-p^{\mathsf{Doc}})(1-\lambda\pi^{\mathsf{Doc}})]}{(1-\gamma)p^{\mathsf{Doc}}\lambda}<1$, where the inequality follows from Assumption 4. Thus $D=0$. ■

Proof of Lemma 2.

According to Lemma 1, $X^{\mathsf{Doc}}_{W^{\mathsf{Doc}}}=0$ and $X^{\mathsf{AI}}_{W^{\mathsf{AI}}}=1$ when $D=0$ and $A=1$. If, in addition, $X^{\mathsf{AI}}_{W^{\mathsf{Doc}}}=1$, the likelihood ratio of the disease, $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{Doc}},W^{\mathsf{Doc}},X^{\mathsf{AI}},W^{\mathsf{AI}})}{\mathbf{P}(Z=0|X^{\mathsf{Doc}},W^{\mathsf{Doc}},X^{\mathsf{AI}},W^{\mathsf{AI}})}$, can be shown as

$$\begin{cases}+\infty, \text{ if } (\max\{X^{\mathsf{Doc}}_L,X^{\mathsf{AI}}_L\},\max\{X^{\mathsf{Doc}}_R,X^{\mathsf{AI}}_R\})=(1,1),\\ \dfrac{\gamma[p^{\mathsf{Doc}}p^{\mathsf{AI}}(1-\lambda)+\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})(1-p^{\mathsf{Doc}}-p^{\mathsf{AI}}+2p^{\mathsf{Doc}}p^{\mathsf{AI}})]}{(1-\gamma)(1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})\lambda}, \text{ otherwise.}\end{cases} \tag{2}$$

Because $\gamma(1-\lambda)\geq(1-\gamma)\lambda$ and $p^{\mathsf{Doc}}p^{\mathsf{AI}}\geq(1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})$, this likelihood ratio is weakly greater than one. Hence, the doctor is persuaded to make $F=1$.

Now, suppose $X^{\mathsf{AI}}_{W^{\mathsf{Doc}}}=0$. The likelihood ratio of the disease is then

$$\frac{\gamma[p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})(1-\lambda)+\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})(p^{\mathsf{Doc}}+p^{\mathsf{AI}}-2p^{\mathsf{Doc}}p^{\mathsf{AI}})]}{(1-\gamma)p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})\lambda}, \tag{3}$$

which is at least one if and only if

$$p^{\mathsf{Doc}}\leq p^{\mathsf{AI}}\cdot\frac{(1-\lambda)/\lambda+(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}{(1-p^{\mathsf{AI}})(1-\gamma)/\gamma+p^{\mathsf{AI}}(1-\lambda)/\lambda+(2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}. \tag{4}$$

Denote the right-hand side as $p_1(\pi^{\mathsf{Doc}})$. We conclude that when $X^{\mathsf{AI}}_{W^{\mathsf{Doc}}}=0$, $F=1$ if and only if $p^{\mathsf{Doc}}\leq p_1(\pi^{\mathsf{Doc}})$. ■

Proof of Corollary 1.

Suppose that the attention difference occurs. Decompose $\mathbf{P}(Z=1|\mathsf{Atten},\mathcal{I}^{\mathsf{Doc}}_{0,1})$ as:

$$\begin{aligned}\mathbf{P}(Z=1|\mathsf{Atten},\mathcal{I}^{\mathsf{Doc}}_{0,1})=&\mathbf{P}(Z=1|\mathcal{E},\mathsf{Atten},\mathcal{I}^{\mathsf{Doc}}_{0,1})\cdot\mathbf{P}(\mathcal{E}|\mathsf{Atten},\mathcal{I}^{\mathsf{Doc}}_{0,1})\\&+\mathbf{P}(Z=1|\mathcal{F},\mathsf{Atten},\mathcal{I}^{\mathsf{Doc}}_{0,1})\cdot\mathbf{P}(\mathcal{F}|\mathsf{Atten},\mathcal{I}^{\mathsf{Doc}}_{0,1}),\end{aligned}$$

where $\mathcal{E}$ denotes the event that $(\max\{X_L^{\mathsf{Doc}}, X_L^{\mathsf{AI}}\}, \max\{X_R^{\mathsf{Doc}}, X_R^{\mathsf{AI}}\}) = (1,1)$, and $\mathcal{F}$ denotes the event that complements $\mathcal{E}$. According to (2), $\mathbf{P}(Z = 1|\mathcal{E}, \mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) = 1$ and $\mathbf{P}(Z = 1|\mathcal{F}, \mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$. Since $\mathbf{P}(\mathcal{E}|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}})$ and $\mathbf{P}(\mathcal{F}|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}})$ are strictly positive, $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) > 1$.

Then, suppose that the comprehension difference occurs. According to (3) and (4), $\mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$ if and only if $p^{\mathsf{Doc}} \leq p_1(\pi^{\mathsf{Doc}})$. ■

Proof of Proposition 1.

Without loss of generality, suppose $W^{\mathsf{Doc}} = L$. There are two cases to consider about the doctor's attention signal: (i) $X^{\mathsf{Doc}} = (0,0)$ and (ii) $X^{\mathsf{Doc}} = (0,1)$.

In (i), $\dfrac{\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}})}{\mathbf{P}(Z = 0|\mathcal{I}_{0,1}^{\mathsf{Doc}})} = \dfrac{\gamma[p^{\mathsf{AI}}(1 - \lambda\pi^{\mathsf{Doc}}) + (1 - p^{\mathsf{AI}})\lambda(1 - \pi^{\mathsf{Doc}})]}{(1 - \gamma)(1 - p^{\mathsf{AI}})\lambda} \geq 1$, where the inequality follows from $\gamma(1 - \lambda\pi^{\mathsf{Doc}}) \geq (1 - \gamma)\lambda$ and $p^{\mathsf{AI}} \geq \frac{1}{2}$. Thus $F = 1$.

In (ii), $\dfrac{\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}})}{\mathbf{P}(Z = 0|\mathcal{I}_{0,1}^{\mathsf{Doc}})} = \dfrac{\gamma[\lambda(1 - \pi^{\mathsf{Doc}})(p^{\mathsf{Doc}}p^{\mathsf{AI}} + 1 - p^{\mathsf{AI}}) + p^{\mathsf{AI}}(1 - p^{\mathsf{Doc}})(1 - \lambda\pi^{\mathsf{Doc}})]}{(1 - \gamma)p^{\mathsf{Doc}}(1 - p^{\mathsf{AI}})\lambda}$.

Because $\dfrac{\partial}{\partial p^{\mathsf{Doc}}}[\lambda(1-\pi^{\mathsf{Doc}})(p^{\mathsf{Doc}}p^{\mathsf{AI}}+1-p^{\mathsf{AI}})+p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})(1-\lambda\pi^{\mathsf{Doc}})] = -(1-\lambda)p^{\mathsf{AI}} \leq 0$, this likelihood ratio decreases in p^{Doc}. In particular, it is at least one if and only if

$$p^{\mathsf{Doc}} \leq \frac{(1 - \lambda)p^{\mathsf{AI}}/\lambda + (1 - \pi^{\mathsf{Doc}})}{(1 - \gamma)(1 - p^{\mathsf{AI}})/\gamma + (1 - \lambda)p^{\mathsf{AI}}/\lambda}.$$

Denote the right-hand side as $p_2(\pi^{\mathsf{Doc}})$, and notice for case (ii) that $F = 1$ if and only if $p^{\mathsf{Doc}} \leq p_2(\pi^{\mathsf{Doc}})$. Then, combining cases (i) and (ii), we conclude that if $p^{\mathsf{Doc}} \leq p_2(\pi^{\mathsf{Doc}})$, the uninterpretable AI persuades the doctor to make $F = 1$.

To show the averaging effect, suppose that $X^{\mathsf{Doc}} = (0,1)$ and $p^{\mathsf{Doc}} = p_1(\pi^{\mathsf{Doc}})$. According to Corollary 1, $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) > \mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) = \frac{1}{2}$. Equation (1) then implies that $\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$, and the inequality is strict as long as $\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \neq 0$. However, according to the definition of $p_2(\pi^{\mathsf{Doc}})$, $\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}}) = \frac{1}{2}$ when $p^{\mathsf{Doc}} = p_2(\pi^{\mathsf{Doc}})$. Since $\dfrac{\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}})}{\mathbf{P}(Z = 0|\mathcal{I}_{0,1}^{\mathsf{Doc}})}$ decreases in p^{Doc}, $p_2(\pi^{\mathsf{Doc}}) \geq p_1(\pi^{\mathsf{Doc}})$. Finally, note that if $\pi^{\mathsf{Doc}} < 1$, $\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) > 0$. The same logic shows that $p_2(\pi^{\mathsf{Doc}}) > p_1(\pi^{\mathsf{Doc}})$ whenever $\pi^{\mathsf{Doc}} < 1$.

To show the attribution effect, again, suppose $X^{\text{Doc}} = (0,1)$. We have

$$\frac{\mathbf{P}(\text{Atten}|\mathcal{I}_{0,1}^{\text{Doc}})}{\mathbf{P}(\text{Comp}|\mathcal{I}_{0,1}^{\text{Doc}})} = \frac{\gamma\lambda(1-\pi^{\text{Doc}})\begin{bmatrix} p^{\text{Doc}}p^{\text{AI}} + (1-p^{\text{Doc}})(1-p^{\text{AI}}) \\ + \pi^{\text{AI}}(p^{\text{Doc}} + p^{\text{AI}} - 2p^{\text{Doc}}p^{\text{AI}}) \end{bmatrix}}{\begin{matrix} \gamma(1-\lambda)p^{\text{AI}}(1-p^{\text{Doc}}) + \lambda(1-\gamma)p^{\text{Doc}}(1-p^{\text{AI}}) \\ + \gamma\lambda(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}})(p^{\text{Doc}} + p^{\text{AI}} - 2p^{\text{Doc}}p^{\text{AI}}) \end{matrix}}, \text{ and}$$

$$\frac{\partial}{\partial\pi^{\text{Doc}}}\frac{\mathbf{P}(\text{Atten}|\mathcal{I}_{0,1}^{\text{Doc}})}{\mathbf{P}(\text{Comp}|\mathcal{I}_{0,1}^{\text{Doc}})} = -\frac{\begin{matrix}[\gamma(1-\lambda)p^{\text{AI}}(1-p^{\text{Doc}}) + \lambda(1-\gamma)p^{\text{Doc}}(1-p^{\text{AI}})]\cdot \\ \gamma\lambda\begin{bmatrix} p^{\text{Doc}}p^{\text{AI}} + (1-p^{\text{Doc}})(1-p^{\text{AI}}) \\ + \pi^{\text{AI}}(p^{\text{Doc}} + p^{\text{AI}} - 2p^{\text{Doc}}p^{\text{AI}}) \end{bmatrix}\end{matrix}}{\begin{bmatrix} \gamma(1-\lambda)p^{\text{AI}}(1-p^{\text{Doc}}) + \lambda(1-\gamma)p^{\text{Doc}}(1-p^{\text{AI}}) \\ + \gamma\lambda(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}})(p^{\text{Doc}} + p^{\text{AI}} - 2p^{\text{Doc}}p^{\text{AI}}) \end{bmatrix}^2} < 0.$$

Since $\mathbf{P}(\text{Atten}|\mathcal{I}_{0,1}^{\text{Doc}}) + \mathbf{P}(\text{Comp}|\mathcal{I}_{0,1}^{\text{Doc}}) = 1$, $\frac{\partial}{\partial\pi^{\text{Doc}}}\frac{\mathbf{P}(\text{Atten}|\mathcal{I}_{0,1}^{\text{Doc}})}{\mathbf{P}(\text{Comp}|\mathcal{I}_{0,1}^{\text{Doc}})} < 0$ implies that $\mathbf{P}(\text{Atten}|\mathcal{I}_{0,1}^{\text{Doc}})$ decreases in π^{Doc}. Further, we can calculate the following derivatives:

$$\frac{\partial p_1(\pi^{\text{Doc}})}{\partial\pi^{\text{Doc}}} = -\frac{(\frac{1-\gamma}{\gamma} + \frac{1-\lambda}{\lambda})p^{\text{AI}}(1-p^{\text{AI}})(1-\pi^{\text{AI}})}{[(1-p^{\text{AI}})(1-\gamma)/\gamma + p^{\text{AI}}(1-\lambda)/\lambda + (2p^{\text{AI}}-1)(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}})]^2}, \quad (5)$$

$$\text{and } \frac{\partial p_2(\pi^{\text{Doc}})}{\partial\pi^{\text{Doc}}} = -\frac{1}{(1-p^{\text{AI}})(1-\gamma)/\gamma + p^{\text{AI}}(1-\lambda)/\lambda}. \quad (6)$$

It is worth noting that $\frac{\partial p_2(\pi^{\text{Doc}})}{\partial\pi^{\text{Doc}}} < 0$. Since $(\frac{1-\gamma}{\gamma} + \frac{1-\lambda}{\lambda})p^{\text{AI}}(1-p^{\text{AI}})(1-\pi^{\text{AI}})$ and $(1-p^{\text{AI}})(1-\gamma)/\gamma + p^{\text{AI}}(1-\lambda)/\lambda$ are both smaller than the denominator of $\frac{\partial p_1(\pi^{\text{Doc}})}{\partial\pi^{\text{Doc}}}$, comparing (5) and (6) shows that $\frac{\partial p_2(\pi^{\text{Doc}})}{\partial\pi^{\text{Doc}}} \leq \frac{\partial p_1(\pi^{\text{Doc}})}{\partial\pi^{\text{Doc}}}$. Therefore, $p_2(\pi^{\text{Doc}}) - p_1(\pi^{\text{Doc}})$ decreases in π^{Doc}, and it increases as π^{Doc} decreases. ■

Proof of Lemma A.1.

According to Lemma 1, $X_{W^{\text{Doc}}}^{\text{Doc}} = 1$ and $X_{W^{\text{AI}}}^{\text{AI}} = 0$ when $D = 1$ and $A = 0$. If, in addition, $W^{\text{AI}} = W^{\text{Doc}}$ or $X^{\text{Doc}} = (1,1)$, then the likelihood ratio of the disease, $\frac{\mathbf{P}(Z=1|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})}{\mathbf{P}(Z=0|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})}$, can be shown to be identical with (2). Since the likelihood ratio is weakly greater than one, the doctor makes $F = 1$.

Now, suppose $W^{\text{AI}} \neq W^{\text{Doc}}$ and $X^{\text{Doc}} \neq (1,1)$. The likelihood ratio of the disease is

then

$$\frac{\gamma[p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})(1-\lambda)+\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})(p^{\mathsf{Doc}}+p^{\mathsf{AI}}-2p^{\mathsf{Doc}}p^{\mathsf{AI}})]}{(1-\gamma)p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})\lambda}, \tag{7}$$

which is smaller than one if and only if

$$p^{\mathsf{Doc}} < p^{\mathsf{AI}} \cdot \frac{(1-\gamma)/\gamma-(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}{p^{\mathsf{AI}}(1-\gamma)/\gamma+(1-p^{\mathsf{AI}})(1-\lambda)/\lambda-(2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}. \tag{8}$$

Denote the right-hand side as $p_3(\pi^{\mathsf{Doc}})$. We conclude that when $W^{\mathsf{AI}} \neq W^{\mathsf{Doc}}$ and $X^{\mathsf{Doc}} \neq (1,1)$, $F=0$ if and only if $p^{\mathsf{Doc}} < p_3(\pi^{\mathsf{Doc}})$. ■

Proof of Corollary A.1.

Suppose that the attention difference occurs. Notice that

$$\mathbf{P}(Z=0|\mathsf{Atten},\mathcal{I}_{1,0}^{\mathsf{Doc}}) = \mathbf{P}(Z=0|\mathcal{F},\mathsf{Atten},\mathcal{I}_{1,0}^{\mathsf{Doc}})\cdot\mathbf{P}(\mathcal{F}|\mathsf{Atten},\mathcal{I}_{1,0}^{\mathsf{Doc}}),$$

where $\mathcal{F}$ denotes the event that $(\max\{X_L^{\mathsf{Doc}},X_L^{\mathsf{AI}}\},\max\{X_R^{\mathsf{Doc}},X_R^{\mathsf{AI}}\}) \neq (1,1)$. According to (2), $\mathbf{P}(Z=0|\mathcal{F},\mathsf{Atten},\mathcal{I}_{1,0}^{\mathsf{Doc}}) \leq \frac{1}{2}$. Since $\mathbf{P}(\mathcal{F}|\mathsf{Atten},\mathcal{I}_{1,0}^{\mathsf{Doc}})$ is strictly smaller than one, $\mathbf{P}(Z=1|\mathsf{Atten},\mathcal{I}_{1,0}^{\mathsf{Doc}}) < \frac{1}{2}$.

Then, suppose that the comprehension difference occurs. According to (7) and (8), $\mathbf{P}(Z=0|\mathsf{Comp},\mathcal{I}_{1,0}^{\mathsf{Doc}}) \geq \frac{1}{2}$ if and only if $X^{\mathsf{Doc}} \neq (1,1)$ and $p^{\mathsf{Doc}} \leq p_3(\pi^{\mathsf{Doc}})$. ■

Proof of Proposition A.1.

If $X^{\mathsf{Doc}}=(1,1)$, it is clear that the doctor does not change her diagnosis. In the following, suppose $W^{\mathsf{Doc}}=L$ and $X^{\mathsf{Doc}}=(1,0)$. The likelihood ratio of the disease is given by

$$\frac{\mathbf{P}(Z=1|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(Z=0|\mathcal{I}_{1,0}^{\mathsf{Doc}})} = \frac{\gamma[\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})(1-\lambda)+p^{\mathsf{Doc}}p^{\mathsf{AI}}(1-\pi^{\mathsf{AI}})(1-\lambda)]}{(1-\gamma)\lambda[p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})+(1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{AI}})]}.$$

Because $\frac{\partial}{\partial p^{\mathsf{Doc}}}[\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})(1-\lambda)+p^{\mathsf{Doc}}p^{\mathsf{AI}}(1-\pi^{\mathsf{AI}})(1-\lambda)] = (1-\lambda)(1-p^{\mathsf{AI}}\pi^{\mathsf{AI}}) \geq 0$, and $1-p^{\mathsf{Doc}}$ is decreasing, this likelihood ratio increases in p^{Doc}. In particular, it is strictly smaller than one if and only if

$$p^{\mathsf{Doc}} < \frac{(1-\gamma)(1-\pi^{\mathsf{AI}}+p^{\mathsf{AI}}\pi^{\mathsf{AI}})/\gamma-(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}{(1-\gamma)(1-\pi^{\mathsf{AI}}+p^{\mathsf{AI}}\pi^{\mathsf{AI}})/\gamma+(1-\lambda)(1-p^{\mathsf{AI}}\pi^{\mathsf{AI}})/\lambda}.$$

Denote the right-hand side as $p_4(\pi^{\mathsf{Doc}})$. We conclude that if $p^{\mathsf{Doc}} < p_4(\pi^{\mathsf{Doc}})$, the

uninterpretable AI persuades the doctor. The averaging effect then follows from this result and Lemma A.1.

To show $p_4(\pi^{\mathsf{Doc}}) \leq p_3(\pi^{\mathsf{Doc}})$, suppose that $p^{\mathsf{Doc}} = p_3(\pi^{\mathsf{Doc}})$. According to Corollary A.1, $\mathbf{P}(Z=0|\mathsf{Atten}, \mathcal{I}_{1,0}^{\mathsf{Doc}}) < \mathbf{P}(Z=0|\mathsf{Comp}, \mathcal{I}_{1,0}^{\mathsf{Doc}}) = \frac{1}{2}$. Then,

$$\begin{aligned}\mathbf{P}(Z=0|\mathcal{I}_{1,0}^{\mathsf{Doc}}) =& \mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}}) \cdot \mathbf{P}(Z=0|\mathsf{Comp}, \mathcal{I}_{1,0}^{\mathsf{Doc}}) \\ &+ \mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}}) \cdot \mathbf{P}(Z=0|\mathsf{Atten}, \mathcal{I}_{1,0}^{\mathsf{Doc}}) \leq \frac{1}{2}.\end{aligned}$$

However, according to the definition $p_4(\pi^{\mathsf{Doc}})$, $\mathbf{P}(Z=0|\mathcal{I}_{1,0}^{\mathsf{Doc}}) = \frac{1}{2}$ when $p^{\mathsf{AI}} = p_4(\pi^{\mathsf{Doc}})$. Since $\dfrac{\mathbf{P}(Z=0|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(Z=1|\mathcal{I}_{1,0}^{\mathsf{Doc}})}$ decreases in p^{Doc}, $p_4(\pi^{\mathsf{Doc}}) \leq p_3(\pi^{\mathsf{Doc}})$.

To show the attribution effect, we first notice that $p_4(\pi^{\mathsf{Doc}})$ may be smaller than $\frac{1}{2}$. Indeed, $p_4(\pi^{\mathsf{Doc}}) \geq \frac{1}{2}$ if and only if

$$\left[\frac{1-\lambda}{\lambda}p^{\mathsf{AI}} - \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}})\right]\pi^{\mathsf{AI}} \geq \frac{1}{\lambda} - \frac{1}{\gamma} + 2(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}}). \tag{9}$$

We must focus on the range of π^{Doc} where (9) holds. We have

$$\frac{\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}})} = \frac{\begin{array}{r}\gamma(1-\lambda)p^{\mathsf{Doc}}(1-p^{\mathsf{AI}}) + \lambda(1-\gamma)p^{\mathsf{AI}}(1-p^{\mathsf{Doc}}) \\ + \gamma\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})(p^{\mathsf{Doc}}+p^{\mathsf{AI}}-2p^{\mathsf{Doc}}p^{\mathsf{AI}})\end{array}}{\begin{array}{r}\gamma(1-\lambda)p^{\mathsf{Doc}}p^{\mathsf{AI}}(1-\pi^{\mathsf{AI}}) + \lambda(1-\gamma)(1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{AI}}) \\ + \gamma\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})[p^{\mathsf{Doc}}p^{\mathsf{AI}} + (1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})]\end{array}}, \text{ and}$$

$$\frac{\partial}{\partial\pi^{\mathsf{Doc}}}\frac{\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}})} = \frac{\gamma\lambda(1-\pi^{\mathsf{AI}})\left[\begin{array}{r}\gamma(1-\lambda)(p^{\mathsf{Doc}})^2p^{\mathsf{AI}}(1-p^{\mathsf{AI}})\pi^{\mathsf{AI}} + \lambda(1-\gamma)(1-p^{\mathsf{Doc}})^2p^{\mathsf{AI}}(1-p^{\mathsf{AI}})\pi^{\mathsf{AI}} \\ + p^{\mathsf{Doc}}(1-p^{\mathsf{Doc}})\left[\begin{array}{r}\lambda(1-\gamma)(1-p^{\mathsf{AI}})^2\pi^{\mathsf{AI}} \\ + \gamma(1-\lambda)(p^{\mathsf{AI}})^2\pi^{\mathsf{AI}} - (\gamma-\lambda)(2p^{\mathsf{AI}}-1)\end{array}\right]\end{array}\right]}{\left[\begin{array}{r}\gamma(1-\lambda)p^{\mathsf{Doc}}p^{\mathsf{AI}}(1-\pi^{\mathsf{AI}}) + \lambda(1-\gamma)(1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{AI}}) \\ + \gamma\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})[p^{\mathsf{Doc}}p^{\mathsf{AI}} + (1-p^{\mathsf{Doc}})(1-p^{\mathsf{AI}})]\end{array}\right]^2}.$$

To show $\frac{\partial}{\partial \pi^{\mathsf{Doc}}} \frac{\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}})} \geq 0$, it suffices to prove

$$\lambda(1-\gamma)(1-p^{\mathsf{AI}})^2\pi^{\mathsf{AI}} + \gamma(1-\lambda)(p^{\mathsf{AI}})^2\pi^{\mathsf{AI}} - (\gamma-\lambda)(2p^{\mathsf{AI}}-1) \geq 0. \tag{10}$$

Note that (9) implies the following lower bound of π^{AI}:

$$\pi^{\mathsf{AI}} \geq \frac{\frac{1}{\lambda} - \frac{1}{\gamma}}{\left[\frac{1-\lambda}{\lambda}p^{\mathsf{AI}} - \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}})\right]}. \tag{11}$$

By substituting (11) into (10), we can get a sufficient condition for (10):

$$\lambda(1-\gamma)(1-p^{\mathsf{AI}})^2 + \gamma(1-\lambda)(p^{\mathsf{AI}})^2 \geq (2p^{\mathsf{AI}}-1)\left[\gamma(1-\lambda)p^{\mathsf{AI}} - \lambda(1-\gamma)(1-p^{\mathsf{AI}})\right].$$

It is easy to verify this inequality by collecting terms. This proves $\frac{\partial}{\partial \pi^{\mathsf{Doc}}} \frac{\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}})} \geq 0$. Since $\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}}) + \mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}}) = 1$, $\frac{\partial}{\partial \pi^{\mathsf{Doc}}} \frac{\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}})}{\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{1,0}^{\mathsf{Doc}})} \geq 0$ implies that $\mathbf{P}(\mathsf{Comp}|\mathcal{I}_{1,0}^{\mathsf{Doc}})$ increases in π^{Doc}.

Further, we calculate the following derivatives:

$$\frac{\partial p_3(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} = \frac{(\frac{1-\gamma}{\gamma} + \frac{1-\lambda}{\lambda})p^{\mathsf{AI}}(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{AI}})}{[p^{\mathsf{AI}}(1-\gamma)/\gamma + (1-p^{\mathsf{AI}})(1-\lambda)/\lambda - (2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})]^2},$$

$$\frac{\partial p_4(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} = \frac{1-\pi^{\mathsf{AI}}}{(1-\gamma)(1-\pi^{\mathsf{AI}}+p^{\mathsf{AI}}\pi^{\mathsf{AI}})/\gamma + (1-\lambda)(1-p^{\mathsf{AI}}\pi^{\mathsf{AI}})/\lambda}.$$

It is worth nothing $\frac{\partial p_4(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} \geq 0$. To show $\frac{\partial p_4(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} \geq \frac{\partial p_3(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}}$, it suffices to show

$$\begin{aligned}(\frac{1}{\gamma} + \frac{1}{\lambda} - 2)p^{\mathsf{AI}}(1-p^{\mathsf{AI}})\left[\frac{1-\gamma}{\gamma}(1-\pi^{\mathsf{AI}}+p^{\mathsf{AI}}\pi^{\mathsf{AI}}) + \frac{1-\lambda}{\lambda}(1-p^{\mathsf{AI}}\pi^{\mathsf{AI}})\right] \\ \leq \left[\frac{1-\gamma}{\gamma}p^{\mathsf{AI}} + \frac{1-\lambda}{\lambda}(1-p^{\mathsf{AI}}) - (2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})\right]^2 =: K^2.\end{aligned} \tag{12}$$

Denote the term in the right-hand side bracket as K. Note that it is increasing as a

function of π^{Doc} and π^{AI}. And recall that by Assumption 4, $\pi^{\text{AI}} \geq \frac{1}{\lambda} - (\frac{1}{\gamma} + \frac{1}{\lambda} - 2)p^{\text{AI}}$. Then, by substituting this inequality and $\pi^{\text{Doc}} = 0$ into K, we obtain one of its lower bounds:

$$K \geq 2\left(\frac{1}{\gamma} + \frac{1}{\lambda} - 2\right) p^{\text{AI}}(1 - p^{\text{AI}}).$$

In particular, $K \geq 0$. This lower bound allows us to get the following sufficient condition for (12):

$$2K \geq \left[\frac{1-\gamma}{\gamma}(1 - \pi^{\text{AI}} + p^{\text{AI}}\pi^{\text{AI}}) + \frac{1-\lambda}{\lambda}(1 - p^{\text{AI}}\pi^{\text{AI}})\right], \text{ or equivalently,}$$

$$\begin{aligned} &p^{\text{AI}}\left[2\frac{1-\gamma}{\gamma} + \frac{1-\lambda}{\lambda}\pi^{\text{AI}} - 2(1 - \pi^{\text{Doc}})(1 - \pi^{\text{AI}})\right] \\ &\quad + (1 - p^{\text{AI}})\left[2\frac{1-\lambda}{\lambda} + \frac{1-\gamma}{\gamma}\pi^{\text{AI}} + 2(1 - \pi^{\text{Doc}})(1 - \pi^{\text{AI}})\right] \geq \frac{1}{\gamma} + \frac{1}{\lambda} - 2. \end{aligned} \tag{13}$$

The left-hand side of (13) can be seen a convex combination of two terms, and each of them is greater than the right-hand side because of (9). This completes the proof of $\frac{\partial p_4(\pi^{\text{Doc}})}{\partial \pi^{\text{Doc}}} \geq \frac{\partial p_3(\pi^{\text{Doc}})}{\partial \pi^{\text{Doc}}}$. Therefore, $p_3(\pi^{\text{Doc}}) - p_4(\pi^{\text{Doc}})$ decreases in π^{Doc}. ■

Comparison of the thresholds for full persuasion:

According to Proposition 1 and Proposition A.1, $p_1(\pi^{\text{Doc}}) \leq p_2(\pi^{\text{Doc}})$ and $p_4(\pi^{\text{Doc}}) \leq p_3(\pi^{\text{Doc}})$. To rank these four thresholds, we only need to compare $p_1(\pi^{\text{Doc}})$ and $p_3(\pi^{\text{Doc}})$. Since $\frac{1-\lambda}{\lambda} \geq \frac{1-\gamma}{\gamma}$ and $2p^{\text{AI}} - 1 \leq 1$ in (4), $p_1(\pi^{\text{Doc}}) \geq p^{\text{AI}}$. In (8), $p_3(\pi^{\text{Doc}}) \leq p^{\text{AI}}$ for the same reason. This proves $p_3(\pi^{\text{Doc}}) \leq p_1(\pi^{\text{Doc}})$. As a result, $p_4(\pi^{\text{Doc}}) \leq p_3(\pi^{\text{Doc}}) \leq p_1(\pi^{\text{Doc}}) \leq p_2(\pi^{\text{Doc}})$. ■

Lemma B.1. *Consider the setting in Section 4, and suppose $p^{\text{Doc}_H} < p_2(\pi^{\text{Doc}})$ and $p^{\text{Doc}_L} > p_3(\pi^{\text{Doc}})$. The behavior of low-type doctors differs between the interpretable and uninterpretable AI only in the following two cases:*

(i) When $D = A = 0$ and $X^{\text{AI}}_{W^{\text{Doc}}} = 1$, low-type doctors with the interpretable AI make $F = 1$, while those with the uninterpretable AI make $F = 0$.

(ii) When $D = 0$, $A = 1$, and $X^{\text{AI}}_{W^{\text{Doc}}} = 0$, low-type doctors with the interpretable AI make $F = 0$, while those with the uninterpretable AI make $F = 1$.

Proof. When doctors only care about their reputation, low-type doctors always mimic high-type doctors. Therefore, we can focus on the behavior of high-type doctors. We consider the equilibrium in which they make efficient diagnoses.

Note that $p_4(\pi^{\text{Doc}}) \leq p_3(\pi^{\text{Doc}}) < p^{\text{Doc}_L} < p^{\text{Doc}_H} < p_2(\pi^{\text{Doc}})$. According to Lemma A.1 and Proposition A.1, when $D = 1$ and $A = 0$, high-type doctors never follow the AI. In contrast, Lemma 2 and Proposition 1 imply that when $D = 0$, $A = 1$, and $X^{\text{AI}}_{W^{\text{Doc}}} = 1$, high-type doctors follow both the interpretable and uninterpretable AI. When $D = 0$, $A = 1$, and $X^{\text{AI}}_{W^{\text{Doc}}} = 0$, high-type doctors only follow the uninterpretable AI.

It remains to compare the interpretable and uninterpretable cases when the doctor and AI agree on the disease diagnosis. There are two such cases: (1) $D = A = 0$ and (2) $D = A = 1$. Note that high-type doctors never change their critical dimension because $p^{\text{Doc}_H} = 1$, and their final diagnosis must depend on the observation in dimension W^{Doc}.

In the first case, high-type doctors do not observe an abnormality in their critical dimension, i.e., $X^{\text{Doc}}_{W^{\text{Doc}}} = 0$. If AI is interpretable and $X^{\text{AI}}_{W^{\text{Doc}}} = 1$, high-type doctors know that there is an abnormality in dimension W^{Doc}. Consequently, the disease is certain, and they make $F = 1$. However, if AI is interpretable and $X^{\text{AI}}_{W^{\text{Doc}}} = 0$, the high-type doctors' belief about the disease is

$$\mathbf{P}(Z = 1 | X^{\text{Doc}}_{W^{\text{Doc}}} = X^{\text{AI}}_{W^{\text{Doc}}} = 0, W = W^{\text{Doc}}] = \frac{\gamma(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}})}{\gamma(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}}) + (1-\gamma)},$$

which is smaller than $\frac{1}{2}$ because $\gamma < \frac{1}{2}$. In this case, high-type doctors make $F = 0$. If AI is uninterpretable, Proposition 1 implies that high-type doctors also make $F = 0$.

In the second case, high-type doctors observe an abnormality in their critical dimension, i.e., $X^{\text{Doc}}_{W^{\text{Doc}}} = 1$. Because the disease is certain, high-type doctors do not change their diagnosis regardless of AI's interpretability. Summarizing this analysis completes the proof. ■

Proof of Proposition 2.

The proof consists of three steps: (1) show that the assumption can be satisfied, (2)

show how uninterpretability changes the final diagnosis of both types of doctors, and (3) show that the change improves the total diagnostic accuracy when the doctor is sufficiently likely to be a low type.

Step 1: Let $p^{\text{Doc}_H} = 1$ and $\pi^{\text{Doc}}, \pi^{\text{AI}}$ be arbitrary numbers in $(0,1)$. Note that $p_2(\pi^{\text{Doc}}) > 1$ if $p^{\text{AI}} > 1 - \frac{\gamma}{1-\gamma}(1-\pi^{\text{Doc}})$. Choose such a value of p^{AI}. Since $p_3(\pi^{\text{Doc}}) < p^{\text{AI}}$ for any $p^{\text{AI}}, \pi^{\text{Doc}}, \pi^{\text{AI}} \in (0,1)$, we can select $p^{\text{Doc}_L} \in (\frac{1}{2}, 1)$ such that $p_3(\pi^{\text{Doc}}) < p^{\text{Doc}_L} < p^{\text{AI}}$. This construction satisfies the assumption that $p^{\text{Doc}_L} < p^{\text{AI}}$, $p^{\text{Doc}_H} < p_2(\pi^{\text{Doc}})$, and $p^{\text{Doc}_L} > p_3(\pi^{\text{Doc}})$.

Step 2: This step is done by Lemma B.1.

Step 3: Following Step 2, we can calculate the changes in the ex-ante diagnostic accuracy of low-type doctors due to uninterpretability. Recall that Lemma B.1 has two parts, (i) and (ii). Denote the change in low-type doctors' diagnostic accuracy as Δ_1 for part (i) and Δ_2 for part (ii). We have

$$
\begin{aligned}
\Delta_1 =& \mathbf{P}(Z=0, D=0, A=0, X^{\text{AI}}_{W^{\text{Doc}}} = 1) - \mathbf{P}(Z=1, D=0, A=0, X^{\text{AI}}_{W^{\text{Doc}}} = 1) \\
=& \pi^{\text{AI}}(1-\pi^{\text{Doc}})[\lambda(1-\gamma)p^{\text{AI}}(1-p^{\text{Doc}_L}) - \gamma(1-\lambda)p^{\text{Doc}_L}(1-p^{\text{AI}})] \\
& - \lambda\gamma(p^{\text{Doc}_L} + p^{\text{AI}} - 2p^{\text{Doc}_L}p^{\text{AI}})\pi^{\text{AI}}(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}}), \text{ and} \\
\Delta_2 =& \mathbf{P}(Z=1, D=0, A=1, X^{\text{AI}}_{W^{\text{Doc}}} = 0] - \mathbf{P}(Z=0, D=0, A=1, X^{\text{AI}}_{W^{\text{Doc}}} = 0] \\
=& \lambda\gamma(p^{\text{Doc}_L} + p^{\text{AI}} - 2p^{\text{Doc}_L}p^{\text{AI}})\pi^{\text{AI}}(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}}) \\
& + \pi^{\text{AI}}[\gamma(1-\lambda)p^{\text{AI}}(1-p^{\text{Doc}_L}) - \lambda(1-\gamma)p^{\text{Doc}_L}(1-p^{\text{AI}})].
\end{aligned}
$$

The summation, $\Delta := \Delta_1 + \Delta_2$, shares the same sign with

$$
\begin{aligned}
& p^{\text{AI}}(1-p^{\text{Doc}_L})[\gamma + \lambda - 2\gamma\lambda - \lambda(1-\gamma)\pi^{\text{Doc}_L}] - p^{\text{Doc}_L}(1-p^{\text{AI}})[\gamma + \lambda - 2\gamma\lambda - \gamma(1-\lambda)\pi^{\text{Doc}_L}] \\
\geq & (p^{\text{AI}} - p^{\text{Doc}_L})[\gamma + \lambda - 2\gamma\lambda - \gamma(1-\lambda)\pi^{\text{Doc}_L}] \quad \text{(Notice that } \lambda(1-\gamma) \leq \gamma(1-\lambda) \text{ because } \gamma \geq \lambda) \\
\geq & (p^{\text{AI}} - p^{\text{Doc}_L})\lambda(1-\gamma) > 0. \quad \text{(Notice that } \pi^{\text{Doc}_L} \leq 1 \text{ and } p^{\text{AI}} > p^{\text{Doc}_L})
\end{aligned}
$$

Let $\overline{\tau} = \frac{\Delta}{1+\Delta}$ and $\tau < \overline{\tau}$. Then, uninterpretability improves the average diagnostic accuracy among doctors by at least $\tau \cdot (-1) + (1-\tau)\Delta > \overline{\tau} \cdot (-1) + (1-\overline{\tau})\Delta = 0$. ■

C AI hallucination

This section generalizes our analysis to the case when AI may hallucinate in drawing its attention signal. Denote as $\phi^{\mathsf{AI}} := \mathbf{P}(X_j^{\mathsf{AI}} = 0 | X_j = 0]$ the probability that the AI correctly observes a normality. Thus, $1 - \phi^{\mathsf{AI}}$ represents the probability of AI hallucination. Focusing on the scenario where the doctor makes a negative initial diagnosis but the AI offers a positive diagnosis, we show that if ϕ^{AI} is sufficiently close to one, our previous results remain the same.

In the following, suppose that $\pi^{\mathsf{Doc}} \in [0, \overline{\pi}]$ for some $\overline{\pi} < 1$, and

$$\phi^{\mathsf{AI}} > \max\left\{\frac{\gamma}{1-\gamma}, 1 - \frac{\gamma\lambda}{1-\gamma}\pi^{\mathsf{AI}}(1-\pi^{\mathsf{AI}}), 1 - \frac{\gamma\lambda}{1-\gamma}\pi^{\mathsf{AI}}(1-\pi^{\mathsf{AI}})(1-\overline{\pi})^2, \right.$$
$$\left. 1 - \frac{\gamma\lambda}{\gamma+\lambda-2\gamma\lambda}\pi^{\mathsf{AI}}(1-\overline{\pi})(2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{AI}})\right\}.$$

We refer to this inequality as that ϕ^{AI} is *sufficiently large*.

Lemma C.1. *The AI makes a positive diagnosis (i.e., $A = 1$) if and only if $X^{\mathsf{AI}}_{W^{\mathsf{AI}}} = 1$.*

Lemma C.1 replicates Lemma 1 by showing that as long as the AI does not hallucinate too often, it still makes a positive diagnosis if and only if an abnormality is observed in its critical dimension.

Proof. Without loss of generality, suppose $W^{\mathsf{AI}} = L$. Similar to the proof of Lemma 1, there are four cases to consider about the AI's attention signal: (i) $X^{\mathsf{AI}} = (1,1)$, (ii) $X^{\mathsf{AI}} = (0,0)$, (iii) $X^{\mathsf{AI}} = (1,0)$, and (iv) $X^{\mathsf{AI}} = (0,1)$.

In (i), $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{AI}}, W^{\mathsf{AI}})}{\mathbf{P}(Z=0|X^{\mathsf{AI}}, W^{\mathsf{AI}})} = \dfrac{\gamma\pi^{\mathsf{AI}}}{(1-\gamma)(1-\phi^{\mathsf{AI}})} \geq 1$, where the inequality follows from the fact that ϕ^{AI} is sufficiently large. Thus $A = 1$.

In (ii), $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{AI}}, W^{\mathsf{AI}})}{\mathbf{P}(Z=0|X^{\mathsf{AI}}, W^{\mathsf{AI}})} = \dfrac{\gamma(1-\pi^{\mathsf{AI}})}{(1-\gamma)\phi^{\mathsf{AI}}} < 1$, where the inequality follows from $\gamma(1-\pi^{\mathsf{AI}}) < 1-\gamma$ and that ϕ^{AI} is sufficiently large. Thus $A = 0$.

In (iii), $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{AI}}, W^{\mathsf{AI}})}{\mathbf{P}(Z=0|X^{\mathsf{AI}}, W^{\mathsf{AI}})} =$

$$\frac{\gamma(1-\lambda)p^{\mathsf{AI}}\pi^{\mathsf{AI}}\phi^{\mathsf{AI}} + \gamma(1-\pi^{\mathsf{AI}})[\lambda p^{\mathsf{AI}}\pi^{\mathsf{AI}} + (1-p^{\mathsf{AI}})(\lambda\pi^{\mathsf{AI}} + (1-\lambda)(1-\phi^{\mathsf{AI}}))]}{\lambda(1-\gamma)(1-p^{\mathsf{AI}})\pi^{\mathsf{AI}}\phi^{\mathsf{AI}} + (1-\gamma)(1-\phi^{\mathsf{AI}})[(1-\lambda)(1-p^{\mathsf{AI}})\phi^{\mathsf{AI}} + p^{\mathsf{AI}}(\lambda(1-\pi^{\mathsf{AI}}) + (1-\lambda)\phi^{\mathsf{AI}})]}.$$

Since $\gamma(1-\lambda)p^{\mathsf{AI}} \geq \lambda(1-\gamma)(1-p^{\mathsf{AI}})$, and the second term in the denominator is strictly smaller than that in the nominator when ϕ^{AI} is sufficiently large, this likelihood ratio is weakly greater than one. Thus $A = 1$.

In (iv), $\dfrac{\mathbf{P}(Z=1|X^{\mathsf{AI}}, W^{\mathsf{AI}})}{\mathbf{P}(Z=0|X^{\mathsf{AI}}, W^{\mathsf{AI}})} =$

$$\frac{\gamma\pi^{\mathsf{AI}}[\lambda p^{\mathsf{AI}}(1-\pi^{\mathsf{AI}}) + (1-p^{\mathsf{AI}})(1-\lambda\pi^{\mathsf{AI}})] + \gamma(1-\lambda)(1-\phi^{\mathsf{AI}})(p^{\mathsf{AI}}-\pi^{\mathsf{AI}})}{\lambda(1-\gamma)p^{\mathsf{AI}}\pi^{\mathsf{AI}} + (1-\gamma)(1-\phi^{\mathsf{AI}})[\lambda(1-p^{\mathsf{AI}}) - \lambda\pi^{\mathsf{AI}} + (1-\lambda)\phi^{\mathsf{AI}}]}.$$

By Assumption 4, the first term is strictly greater in the denominator than in the nominator. Since ϕ^{AI} is sufficiently large, the second term is also strictly greater in the denominator. Thus the above term is strictly smaller than one, and $A = 0$. ■

Next, we study AI persuasion when AI is interpretable. We follow Section 3.1 to define the attention difference and the comprehension difference between the doctor and AI.

Lemma C.2. *Suppose that the AI is interpretable and the diagnostic disagreement is given by $D = 0$ and $A = 1$. Then, the following hold:*

(i) When the attention difference occurs, the AI persuades the doctor to change her diagnosis from $D = 0$ to $F = 1$ regardless of her skill $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$.

(ii) When the comprehension difference occurs and $X^{\mathsf{Doc}} = (0,0)$, the AI persuades the doctor if and only if her comprehension skill, p^{Doc}, is weakly below a threshold $p_1'(\pi^{\mathsf{Doc}})$.

(iii) When comprehension difference occurs and $X^{\mathsf{Doc}} \neq (0,0)$, the AI persuades the doctor if and only if her comprehension skill, p^{Doc}, is weakly below a threshold $p_1''(\pi^{\mathsf{Doc}})$. In particular, $p_1'(\pi^{\mathsf{Doc}}) \leq p_1''(\pi^{\mathsf{Doc}})$.

Lemma C.2 is similar to Lemma 2 in showing that the attention difference is persuasive regardless of the doctor's skill, and the comprehension difference is persuasive only if the doctor's comprehension skill is sufficiently low. The only difference from Lemma 2 is that now the persuasiveness of the comprehension difference also depends on what the doctor observes. If the doctor has observed an abnormality, it is easier for the AI to persuade the doctor to make a positive final diagnosis.

Proof. Without loss of generality, suppose $W^{\text{Doc}} = L$. According to Lemma 1 and Lemma C.1, $X_L^{\text{Doc}} = 0$ and $X_{W^{\text{AI}}}^{\text{AI}} = 1$ when $D = 0$ and $A = 1$. If, in addition, $X_L^{\text{AI}} = 1$, we get six cases: (i) $X^{\text{Doc}} = (0,0)$ and $X^{\text{AI}} = (1,0)$, (ii) $X^{\text{Doc}} = (0,0)$, $X^{\text{AI}} = (1,1)$, and $W^{\text{AI}} = L$, (iii) $X^{\text{Doc}} = (0,0)$, $X^{\text{AI}} = (1,1)$, and $W^{\text{AI}} = R$, (iv) $X^{\text{Doc}} = (0,1)$ and $X^{\text{AI}} = (1,0)$, (v) $X^{\text{Doc}} = (0,1)$, $X^{\text{AI}} = (1,1)$, and $W^{\text{AI}} = L$, and (vi) $X^{\text{Doc}} = (0,1)$, $X^{\text{AI}} = (1,1)$, and $W^{\text{AI}} = R$.

In (i), $\dfrac{\mathbf{P}(Z=1|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})}{\mathbf{P}(Z=0|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})} =$

$$\frac{\begin{array}{l}\gamma(1-\lambda)p^{\text{Doc}}p^{\text{AI}}\pi^{\text{AI}}(1-\pi^{\text{Doc}})\phi^{\text{AI}} \\ \quad + \gamma(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}})\left[\begin{array}{l}\lambda p^{\text{Doc}}p^{\text{AI}}\pi^{\text{AI}}(1-\pi^{\text{Doc}}) \\ \quad + (1-p^{\text{Doc}})(1-p^{\text{AI}})(\lambda\pi^{\text{AI}}(1-\pi^{\text{Doc}}) + (1-\lambda)(1-\phi^{\text{AI}}))\end{array}\right]\end{array}}{\begin{array}{l}\lambda(1-\gamma)(1-p^{\text{Doc}})(1-p^{\text{AI}})\pi^{\text{AI}}(1-\pi^{\text{Doc}})\phi^{\text{AI}} \\ \quad + (1-\gamma)(1-\phi^{\text{AI}})\left[\begin{array}{l}(1-\lambda)(1-p^{\text{Doc}})(1-p^{\text{AI}})\phi^{\text{AI}} \\ \quad + p^{\text{Doc}}p^{\text{AI}}(\lambda(1-\pi^{\text{Doc}})(1-\pi^{\text{AI}}) + (1-\lambda)\phi^{\text{AI}})\end{array}\right]\end{array}}.$$

This likelihood ratio is weakly greater than one because $\gamma(1-\lambda)p^{\text{Doc}}p^{\text{AI}} \geq \lambda(1-\gamma)(1-p^{\text{Doc}})(1-p^{\text{AI}})$, and the second term in the denominator is strictly smaller than that in the nominator when ϕ^{AI} is sufficiently large. Thus $F = 1$.

In both (ii) and (iii), $\dfrac{\mathbf{P}(Z=1|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})}{\mathbf{P}(Z=0|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})} = \dfrac{\gamma\pi^{\text{AI}}(1-\pi^{\text{Doc}})}{(1-\gamma)(1-\phi^{\text{AI}})} > 1$, where the strict inequality follows as ϕ^{AI} is sufficiently large. Thus $F = 1$.

In both (iv) and (v), $\dfrac{\mathbf{P}(Z=1|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})}{\mathbf{P}(Z=0|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})} =$

$$\frac{\gamma[\lambda p^{\text{Doc}}p^{\text{AI}}\pi^{\text{AI}}(1-\pi^{\text{Doc}}) + (1-p^{\text{Doc}})(1-p^{\text{AI}})(\lambda\pi^{\text{AI}}(1-\pi^{\text{Doc}}) + (1-\lambda)(1-\phi^{\text{AI}}))]}{\lambda(1-\gamma)p^{\text{Doc}}p^{\text{AI}}(1-\phi^{\text{AI}})},$$

which is weakly greater than one because $\gamma\pi^{\text{AI}}(1-\pi^{\text{Doc}}) \geq (1-\gamma)(1-\phi^{\text{AI}})$ when ϕ^{AI} is sufficiently large. Thus $F = 1$.

In (vi), $\dfrac{\mathbf{P}(Z=1|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})}{\mathbf{P}(Z=0|X^{\text{Doc}}, W^{\text{Doc}}, X^{\text{AI}}, W^{\text{AI}})} =$

$$\frac{\gamma[\lambda p^{\text{Doc}}(1-p^{\text{AI}})\pi^{\text{AI}}(1-\pi^{\text{Doc}}) + p^{\text{AI}}(1-p^{\text{Doc}})(\lambda\pi^{\text{AI}}(1-\pi^{\text{Doc}}) + (1-\lambda)(1-\phi^{\text{AI}}))]}{\lambda(1-\gamma)p^{\text{Doc}}(1-p^{\text{AI}})(1-\phi^{\text{AI}})},$$

which is weakly greater than one because $\gamma\pi^{\text{AI}}(1-\pi^{\text{Doc}}) \geq (1-\gamma)(1-\phi^{\text{AI}})$ when ϕ^{AI}

is sufficiently large. Thus $F = 1$. Summarizing these six cases shows that the attention difference is always persuasive.

Now, consider the comprehension difference, i.e., $X^{\mathsf{AI}} = (0,1)$ and $W^{\mathsf{AI}} = R$. There are two cases: (i) $X^{\mathsf{Doc}} = (0,0)$ and (ii) $X^{\mathsf{Doc}} = (0,1)$. In (i), the likelihood ratio of the disease is

$$\frac{\gamma\begin{bmatrix} p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})(\lambda\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}})+(1-\lambda)(1-\phi^{\mathsf{AI}})) \\ p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}})(\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+(1-\lambda)\phi^{\mathsf{AI}}) \end{bmatrix}}{(1-\gamma)\begin{bmatrix} p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})\phi^{\mathsf{AI}}(\lambda\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}})+(1-\lambda)(1-\phi^{\mathsf{AI}})) \\ p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})(1-\phi^{\mathsf{AI}})(\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+(1-\lambda)\phi^{\mathsf{AI}}) \end{bmatrix}},$$

which is at least one if and only if

$$p^{\mathsf{Doc}} \leq p^{\mathsf{AI}} \cdot \frac{\begin{pmatrix} \gamma\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}}) \\ -(1-\gamma)(1-\phi^{\mathsf{AI}}) \end{pmatrix}(\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+(1-\lambda)\phi^{\mathsf{AI}})}{\begin{bmatrix} p^{\mathsf{AI}}\begin{pmatrix} \gamma\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}}) \\ -(1-\gamma)(1-\phi^{\mathsf{AI}}) \end{pmatrix}(\lambda(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+(1-\lambda)\phi^{\mathsf{AI}}) \\ +(1-p^{\mathsf{AI}})\begin{pmatrix} (1-\gamma)\phi^{\mathsf{AI}} \\ -\gamma(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}}) \end{pmatrix}(\lambda\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}})+(1-\lambda)(1-\phi^{\mathsf{AI}})) \end{bmatrix}}. \tag{14}$$

Denote the right-hand side as $p_1'(\pi^{\mathsf{Doc}})$. Thus for (i), $F = 1$ if and only if $p^{\mathsf{Doc}} \leq p_1'(\pi^{\mathsf{Doc}})$.

In (ii), the likelihood ratio of the disease is

$$\frac{\gamma\begin{bmatrix} \lambda p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}}) \\ +\lambda p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})+(1-\lambda)p^{\mathsf{AI}}(1-p^{\mathsf{Doc}})\phi^{\mathsf{AI}} \end{bmatrix}}{\lambda(1-\gamma)p^{\mathsf{Doc}}(1-p^{\mathsf{AI}})\phi^{\mathsf{AI}}},$$

which is at least one if and only if

$$p^{\mathsf{Doc}} \leq p^{\mathsf{AI}} \cdot \frac{\frac{1-\lambda}{\lambda}\phi^{\mathsf{AI}} + (1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}{\frac{1-\lambda}{\lambda}p^{\mathsf{AI}}\phi^{\mathsf{AI}} + \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}})\phi^{\mathsf{AI}} + (2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})}. \tag{15}$$

Denote the right-hand side as $p_1''(\pi^{\mathsf{Doc}})$. Thus for (ii), $F = 1$ if and only if $p^{\mathsf{Doc}} \leq p_1''(\pi^{\mathsf{Doc}})$. A direct comparison between (14) and (15) reveals that $p_1'(p^{\mathsf{Doc}}) \leq p_1''(\pi^{\mathsf{Doc}})$. ■

As a corollary of Lemma C.2, the following result replicates Corollary 1.

Corollary C.1. *The AI is more persuasive with the attention difference than the comprehension difference:* $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) > \frac{1}{2}$ *for any* $(\pi^{\mathsf{Doc}}, p^{\mathsf{Doc}})$*, but* $\mathbf{P}(Z = 1|\mathsf{Comp}, X^{\mathsf{Doc}} = (0,0), \mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$ *only if* $p^{\mathsf{Doc}} \leq p_1'(\pi^{\mathsf{Doc}})$*, and* $\mathbf{P}(Z = 1|\mathsf{Comp}, X_{-W^{\mathsf{Doc}}}^{\mathsf{Doc}} = 1, \mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$ *only if* $p^{\mathsf{Doc}} \leq p_1''(\pi^{\mathsf{Doc}})$*.*

Finally, we study AI persuasion when AI is uninterpretable. The following proposition replicates Proposition 1.

Proposition C.1. *Suppose that the AI is uninterpretable and the diagnostic disagreement is given by* $D = 0$ *and* $A = 1$*. Then, the following hold:*

(i) **Threshold for persuasion:** *There exists a threshold,* $p_2'(\pi^{\mathsf{Doc}})$*, such that the AI persuades the doctor whenever* $p^{\mathsf{Doc}} \leq p_2'(\pi^{\mathsf{Doc}})$*.*

(ii) **Averaging effect:** $p_2'(\pi^{\mathsf{Doc}}) \geq p_1''(\pi^{\mathsf{Doc}})$*, and the inequality is strict if* $\pi^{\mathsf{Doc}} < 1$*.*

(iii) **Attribution effect:** *As* π^{Doc} *decreases,* $p_2'(\pi^{\mathsf{Doc}}) - p_1''(\pi^{\mathsf{Doc}})$ *increases.*

Proof. Without loss of generality, suppose $W^{\mathsf{Doc}} = L$. There are two cases to consider about the doctor's attention signal: (i) $X^{\mathsf{Doc}} = (0,0)$ and (ii) $X^{\mathsf{Doc}} = (0,1)$.

In (i), $\dfrac{\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}})}{\mathbf{P}(Z = 0|\mathcal{I}_{0,1}^{\mathsf{Doc}})} =$

$$\frac{\gamma(1-\pi^{\mathsf{Doc}})[p^{\mathsf{AI}}\pi^{\mathsf{AI}}(1-\lambda\pi^{\mathsf{Doc}}) + (1-p^{\mathsf{AI}})(\lambda(1-\pi^{\mathsf{Doc}})\pi^{\mathsf{AI}} + (1-\lambda)(1-\phi^{\mathsf{AI}}))]}{\lambda(1-\gamma)(1-p^{\mathsf{AI}})\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}}) + (1-\gamma)(1-\phi^{\mathsf{AI}})[p^{\mathsf{AI}}(1-\lambda\pi^{\mathsf{Doc}}) + (1-\lambda)(1-p^{\mathsf{AI}})]}.$$

Since $\gamma[p^{\mathsf{AI}}(1-\lambda\pi^{\mathsf{Doc}}) + \lambda(1-p^{\mathsf{AI}})] \geq \lambda(1-\gamma)(1-p^{\mathsf{AI}})$, the difference between the nominator and the first term in the denominator is strictly positive. Further, this difference is strictly greater than the second term in the denominator when ϕ^{AI} is sufficiently large. Thus, the likelihood ratio is weakly greater than one, and $F = 1$.

In (ii), $\dfrac{\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}})}{\mathbf{P}(Z = 0|\mathcal{I}_{0,1}^{\mathsf{Doc}})} =$

$$\frac{\gamma\left[\lambda\pi^{\mathsf{AI}}(1-\pi^{\mathsf{Doc}}) + (1-\lambda)(1-p^{\mathsf{Doc}})(p^{\mathsf{AI}}\pi^{\mathsf{AI}} + (1-p^{\mathsf{AI}})(1-\phi^{\mathsf{AI}}))\right]}{\lambda(1-\gamma)p^{\mathsf{Doc}}\left[p^{\mathsf{AI}}(1-\phi^{\mathsf{AI}}) + (1-p^{\mathsf{AI}})\pi^{\mathsf{AI}}\right]}.$$

The likelihood ratio is at least one if and only if

$$p^{\mathsf{Doc}} \leq \frac{\frac{1-\phi^{\mathsf{AI}}}{\pi^{\mathsf{AI}}}\frac{1-\lambda}{\lambda}(1-p^{\mathsf{AI}}) + \frac{1-\lambda}{\lambda}p^{\mathsf{AI}} + (1-\pi^{\mathsf{Doc}})}{\frac{1-\phi^{\mathsf{AI}}}{\pi^{\mathsf{AI}}}\left[\frac{1-\gamma}{\gamma}p^{\mathsf{AI}} + \frac{1-\lambda}{\lambda}(1-p^{\mathsf{AI}})\right] + \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}}) + \frac{1-\lambda}{\lambda}p^{\mathsf{AI}}}.$$

Denote the right-hand side as $p_2'(\pi^{\mathsf{Doc}})$, and notice for case (ii) that $F = 1$ if and only if $p^{\mathsf{Doc}} \leq p_2'(\pi^{\mathsf{Doc}})$. Then, combining cases (i) and (ii), we conclude that if $p^{\mathsf{Doc}} \leq p_2'(\pi^{\mathsf{Doc}})$, the uninterpretable AI persuades the doctor to make $F = 1$.

To show the averaging effect, suppose that $X^{\mathsf{Doc}} = (0,1)$ and $p^{\mathsf{Doc}} = p_1''(\pi^{\mathsf{Doc}})$. According to Corollary C.1, $\mathbf{P}(Z = 1|\mathsf{Atten}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) > \mathbf{P}(Z = 1|\mathsf{Comp}, \mathcal{I}_{0,1}^{\mathsf{Doc}}) = \frac{1}{2}$. Equation (1) then implies that $\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \geq \frac{1}{2}$, and the inequality is strict as long as $\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) \neq 0$. However, according to the definition of $p_2'(\pi^{\mathsf{Doc}})$, $\mathbf{P}(Z = 1|\mathcal{I}_{0,1}^{\mathsf{Doc}}) = \frac{1}{2}$ when $p^{\mathsf{Doc}} = p_2'(\pi^{\mathsf{Doc}})$. Since $\dfrac{\mathbf{P}(Z=1|\mathcal{I}_{0,1}^{\mathsf{Doc}})}{\mathbf{P}(Z=0|\mathcal{I}_{0,1}^{\mathsf{Doc}})}$ decreases in p^{Doc}, $p_2'(\pi^{\mathsf{Doc}}) \geq p_1''(\pi^{\mathsf{Doc}})$. Finally, note that if $\pi^{\mathsf{Doc}} < 1$, $\mathbf{P}(\mathsf{Atten}|\mathcal{I}_{0,1}^{\mathsf{Doc}}) > 0$. The same logic shows that $p_2'(\pi^{\mathsf{Doc}}) > p_1''(\pi^{\mathsf{Doc}})$ whenever $\pi^{\mathsf{Doc}} < 1$.

To show the attribution effect, we can calculate the following derivatives:

$$\frac{\partial p_1''(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} = -\frac{\left(\frac{1-\gamma}{\gamma} + \frac{1-\lambda}{\lambda}\right)p^{\mathsf{AI}}(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{AI}})\phi^{\mathsf{AI}}}{\left[\frac{1-\lambda}{\lambda}p^{\mathsf{AI}}\phi^{\mathsf{AI}} + \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}})\phi^{\mathsf{AI}} + (2p^{\mathsf{AI}}-1)(1-\pi^{\mathsf{Doc}})(1-\pi^{\mathsf{AI}})\right]^2}, \quad (16)$$

$$\text{and } \frac{\partial p_2'(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} = -\frac{1}{\frac{1-\phi^{\mathsf{AI}}}{\pi^{\mathsf{AI}}}\left[\frac{1-\gamma}{\gamma}p^{\mathsf{AI}} + \frac{1-\lambda}{\lambda}(1-p^{\mathsf{AI}})\right] + \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}}) + \frac{1-\lambda}{\lambda}p^{\mathsf{AI}}}. \quad (17)$$

It is worth noting that $\dfrac{\partial p_2'(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} < 0$. When ϕ^{AI} is sufficiently large, both $(\frac{1-\gamma}{\gamma} + \frac{1-\lambda}{\lambda})p^{\mathsf{AI}}(1-p^{\mathsf{AI}})(1-\pi^{\mathsf{AI}})\phi^{\mathsf{AI}}$ and $\frac{1-\phi^{\mathsf{AI}}}{\pi^{\mathsf{AI}}}\left[\frac{1-\gamma}{\gamma}p^{\mathsf{AI}} + \frac{1-\lambda}{\lambda}(1-p^{\mathsf{AI}})\right] + \frac{1-\gamma}{\gamma}(1-p^{\mathsf{AI}}) + \frac{1-\lambda}{\lambda}p^{\mathsf{AI}}$ are smaller than the denominator of $\dfrac{\partial p_1'(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}}$. Then, comparing (16) and (17) shows that $\dfrac{\partial p_2'(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}} \leq \dfrac{\partial p_1''(\pi^{\mathsf{Doc}})}{\partial \pi^{\mathsf{Doc}}}$. Therefore, $p_2'(\pi^{\mathsf{Doc}}) - p_1''(\pi^{\mathsf{Doc}})$ decreases in π^{Doc}, and it increases as π^{Doc} decreases. ■

D Free-riding on AI information

In addition to the model described in Section 2, suppose that the doctor, after she receives the AI information, can incur a cost $c > 0$ to draw an extra attention signal, denoted by X^{E}. The signal shares the same structure with X^{Doc} and is independent of X^{Doc} conditional on the abnormality status X_L and X_R. We provide below an example where AI's uninterpretability can improve diagnostic accuracy. Notice that while the example assumes $\pi^{\mathsf{AI}} = 0$, it is easy to extend the example to accommodate cases when $\pi^{\mathsf{AI}} > 0$ is sufficiently small. We make this assumption only to simplify our analysis.

Example D.1. *Suppose $p^{\mathsf{AI}} < p_1(\pi^{\mathsf{Doc}})$ and $\pi^{\mathsf{AI}} = 0$. There exist $c_1, c_2 > 0$ such that if $c \in (c_1, c_2)$, the following hold:*

(i) The doctor is more likely to draw the extra attention signal when AI is uninterpretable than when it is interpretable.

(ii) The accuracy of the doctor's final diagnosis is higher with uninterpretable AI than with interpretable AI.

Proof. Because the extra attention signal never negates the observation of abnormalities, drawing it will not change the doctor's final diagnosis if without it, the doctor would have made $F = 1$. Moreover, Lemma 2 and Proposition A.1 show that when $p^{\mathsf{AI}} < p_1(\pi^{\mathsf{Doc}}) \leq p_3(\pi^{\mathsf{Doc}})$, $D = 1$ implies $F = 1$. Hence, we can focus on the case when $D = 0$. Because $\pi^{\mathsf{AI}} = 0$, $X^{\mathsf{AI}} = (0, 0)$ and $A = 0$.

Without loss of generality, suppose $W^{\mathsf{Doc}} = L$. The original signals, X^{Doc} and W^{AI}, have four cases: (1) $X^{\mathsf{Doc}} = (0, 0)$ and $W^{\mathsf{AI}} = L$, (2) $X^{\mathsf{Doc}} = (0, 0)$ and $W^{\mathsf{AI}} = R$, (3) $X^{\mathsf{Doc}} = (0, 1)$ and $W^{\mathsf{AI}} = L$, and (4) $X^{\mathsf{Doc}} = (0, 1)$ and $W^{\mathsf{AI}} = R$. In each case, the doctor may change her diagnosis only if she draws another attention signal that shows $X_{W^{\mathsf{Doc}}} = 1$.

For $i \in \{1, 2, 3, 4\}$, denote the original signal profile as $\mathcal{S}_i := (X^{\mathsf{Doc}}, W^{\mathsf{Doc}}, X^{\mathsf{AI}}, W^{\mathsf{AI}})$, the extra attention signal as X^{E}, and the expected increase in diagnostic accuracy from

this extra attention signal as Δ_i. For $i = 1$, we can calculate Δ_1 as follows:

$$\begin{aligned}\Delta_1 =&\mathbf{P}[Z = 1, X_L^{\mathsf{E}} = 1|\mathcal{S}_i] - \mathbf{P}[Z = 0, X_L^{\mathsf{E}} = 1|\mathcal{S}_i] \\ =&\pi^{\mathsf{Doc}}(1 - \pi^{\mathsf{Doc}}) \cdot \frac{\gamma[\mu(1 - \lambda\pi^{\mathsf{Doc}}) + (1 - \mu)\lambda(1 - \pi^{\mathsf{Doc}})] - (1 - \gamma)(1 - \mu)\lambda}{\gamma(1 - \pi^{\mathsf{Doc}})(1 - \lambda\pi^{\mathsf{Doc}}) + (1 - \gamma)(1 - \lambda\pi^{\mathsf{Doc}})},\end{aligned}$$

where $\mu = \dfrac{p^{\mathsf{Doc}}p^{\mathsf{AI}}}{p^{\mathsf{Doc}}p^{\mathsf{AI}} + (1 - p^{\mathsf{Doc}})(1 - p^{\mathsf{AI}})}$. Similarly,

$$\Delta_2 = \pi^{\mathsf{Doc}}(1 - \pi^{\mathsf{Doc}}) \cdot \frac{\gamma[\eta(1 - \lambda\pi^{\mathsf{Doc}}) + (1 - \eta)\lambda(1 - \pi^{\mathsf{Doc}})] - (1 - \gamma)(1 - \eta)\lambda}{\gamma(1 - \pi^{\mathsf{Doc}})(1 - \lambda\pi^{\mathsf{Doc}}) + (1 - \gamma)(1 - \lambda\pi^{\mathsf{Doc}})},$$

where $\eta = \dfrac{p^{\mathsf{Doc}}(1 - p^{\mathsf{AI}})}{p^{\mathsf{Doc}}(1 - p^{\mathsf{AI}}) + p^{\mathsf{AI}}(1 - p^{\mathsf{Doc}})}$. Notice that $\eta > \dfrac{1}{2}$ as $p^{\mathsf{AI}} < p_1(\pi^{\mathsf{Doc}}) \leq p^{\mathsf{Doc}}$. Because $\mu > \eta$, $\Delta_1 > \Delta_2$. In addition, because

$$\begin{aligned}\gamma[\eta(1 - \lambda\pi^{\mathsf{Doc}}) + (1 - \eta)\lambda(1 - \pi^{\mathsf{Doc}})] >&\gamma(1 - \eta)[1 + \lambda - 2\lambda\pi^{\mathsf{Doc}}] \quad (\eta > 1 - \eta) \\ \geq&\gamma(1 - \eta)(1 - \lambda) \quad (\pi^{\mathsf{Doc}} \leq 1) \\ \geq&(1 - \gamma)(1 - \eta)\lambda, \quad (\gamma \geq \lambda)\end{aligned}$$

we know that $\Delta_2 > 0$. In a similar way, we can show that $\Delta_3 > \Delta_4 > 0$. Therefore, it is optimal for the doctor to make $F = 1$ whenever $X_{W^{\mathsf{Doc}}}^{\mathsf{E}} = 1$.

Notice the following facts about Δ_i and the doctor's behavior. When AI is interpretable, the doctor will draw the extra attention signal given case (i) if and only if $c \leq \Delta_i$. In contrast, when AI is uninterpretable, the doctor cannot distinguish between cases (1) and (2) and between cases (3) and (4). Given $X^{\mathsf{Doc}} = (0, 0)$, the doctor will draw the extra attention signal if and only if $c \leq \mathbf{P}[\mathcal{S}_1|X^{\mathsf{Doc}} = (0, 0), W^{\mathsf{Doc}} = L] \cdot \Delta_1 + \mathbf{P}[\mathcal{S}_2|X^{\mathsf{Doc}} = (0, 0), W^{\mathsf{Doc}} = L] \cdot \Delta_2$; and given $X^{\mathsf{Doc}} = (0, 1)$, she will draw the extra attention signal if and only if $c \leq \mathbf{P}[\mathcal{S}_3|X^{\mathsf{Doc}} = (0, 1), W^{\mathsf{Doc}} = L] \cdot \Delta_3 + \mathbf{P}[\mathcal{S}_4|X^{\mathsf{Doc}} = (0, 1), W^{\mathsf{Doc}} = L] \cdot \Delta_4$.

Now we construct c_1 and c_2, and let $c \in (c_1, c_2)$. If $\Delta_2 > \Delta_4$, let $c_1 = \Delta_4$ and $c_2 = \min\{\Delta_2, \mathbf{P}[\mathcal{S}_3|X^{\mathsf{Doc}} = (0, 1), W^{\mathsf{Doc}} = L] \cdot \Delta_3 + \mathbf{P}[\mathcal{S}_4|X^{\mathsf{Doc}} = (0, 1), W^{\mathsf{Doc}} = L] \cdot \Delta_4\}$. Then, when AI is uninterpretable, it is always optimal for the doctor to draw the extra attention signal. In contrast, when AI is interpretable, this is optimal only in cases (1),

(2), and (3).

If $\Delta_2 = \Delta_4$, let $c_1 = \Delta_2$ and $c_2 = \min\{\mathbf{P}[\mathcal{S}_1|X^{\mathsf{Doc}} = (0,0), W^{\mathsf{Doc}} = L] \cdot \Delta_1 + \mathbf{P}[\mathcal{S}_2|X^{\mathsf{Doc}} = (0,0), W^{\mathsf{Doc}} = L] \cdot \Delta_2, \mathbf{P}[\mathcal{S}_3|X^{\mathsf{Doc}} = (0,1), W^{\mathsf{Doc}} = L] \cdot \Delta_3 + \mathbf{P}[\mathcal{S}_4|X^{\mathsf{Doc}} = (0,1), W^{\mathsf{Doc}} = L] \cdot \Delta_4\}$. Then, when AI is uninterpretable, it is always optimal for the doctor to draw the extra attention signal. In contrast, when AI is interpretable, this is optimal only in cases (1) and (3).

If $\Delta_4 > \Delta_2$, let $c_1 = \Delta_2$ and $c_2 = \min\{\mathbf{P}[\mathcal{S}_1|X^{\mathsf{Doc}} = (0,0), W^{\mathsf{Doc}} = L] \cdot \Delta_1 + \mathbf{P}[\mathcal{S}_2|X^{\mathsf{Doc}} = (0,0), W^{\mathsf{Doc}} = L] \cdot \Delta_2, \Delta_4\}$. Then, when AI is uninterpretable, it is always optimal for the doctor to draw the extra attention signal. In contrast, when AI is interpretable, this is optimal only in cases (1), (3), and (4).

The construction in the previous three paragraphs shows that the doctor is more likely to draw the extra attention signal when AI is uninterpretable than when it is interpretable. Then, because $\Delta_i > 0$ for any i indicates that drawing the extra attention signal is always beneficial regarding diagnostic accuracy, we conclude that the accuracy of the doctor's final diagnosis is higher with uninterpretable AI than with interpretable AI. ■