AI and decision-making: a Principal-Agent Perspective

Luc Fortin-Tyson*[1]

[1]CRED, Université Panthéon-Assas

March 15, 2024

Abstract

I analyze optimal incentive contracts characterized by asymmetries of information regarding the kind of AI technology used. Indeed, the most explainable AIs are also the least accurate. However, explainability allows to rule out AI errors more easily through an ex-post effort, eventually avoiding high level of losses. I suppose that a decision is delegated by a principal to an agent using an AI to perform it. The type of AI is selected by the agent. Moreover, there is an asymmetry of information regarding the use of AI technology. I analyze the contracts that can be settled between the different parties. I thus look at the potential inefficiencies that can be created by the use of AIs. Indeed, incentivizing the agent to use the correct type of AI implies that the agent extracts a rent. the principal can reduce the level of rent, but will damage the social welfare.

Keywords:

Artificial Intelligence, Principal Agent, Liability, Explainability, Accuracy

JEL: L50, L24, K24

Acknowledgments The author thanks Alexandre Mayol, Chloé Le Coq, and Elena Dumitrescu for their useful comments. The author also thanks his PhD companions Illan Barriola, Florian Thierry-Coudon, and Charlie Noujarret for their valuable feedback. All errors are mine.

*luc.fortin@u-paris2.fr

1 Introduction

AI is extensively used for decision-making and is expected to gain even more importance in the future (Davenport and Harris, 2017). In many settings, incentive conflicts arise between those building the prediction tools and the entities tasked with overseeing their use (Blattner, Nelson, and Spiess, 2023). Indeed, the word "AI" encompasses a large variety of methods, associated with different characteristics. In particular, the paper focuses on the various degrees of **explainability** of AI methods. Indeed, there is a trade-off between the explainability of AI and its accuracy (Doshi-Velez and Kim, 2017). Indeed, the most unexplainable AI models are also the less explainable ones. The question of the explainability of the AIs has been a vibrant topic in the IT field for a few years. Indeed, the explainability of the AI is important to secure the AI and avoid bad outcomes due to wrong AI decisions (Beaudouin et al., 2020).

In this paper, I analyze optimal incentive contracts characterized by asymmetric information regarding the type of AI technology used. This paper is linked to the literature dedicated to principal-agent models (see Holmstrom, 1979, Aghion and Tirole, 1994 or Boyer and Laffont, 1997 for examples). My contribution to the current literature lies in the assumption that agents select endogenously a type of AI technology that shapes the outcomes obtained by the principal. Indeed, AI systems can be complicated enough such that an uninformed principal cannot asses the type of technology used (Blattner, Nelson, and Spiess, 2023). Moreover, AI technology can be considered as trade secrets implying they cannot be examined by the public. For instance, courts denied access to predictive justice algorithms, such as the COMPAS algorithm (Washington, 2018), arguing trade secret clauses, even though it has been established that these algorithms discriminate against sensible groups (Angwin et al., 2022). In addition to the choice of technology, the agent also has private information about the state of the world. In other words, the consequences of the success and errors of the AI are known by the agent, but not by the principal. The choice of the "good type" of AI by the agent crucially depends on the state of the world. Different types of AI, with different characteristics, can be used by the agent. When potential losses associated with a decision are high, one prefers to use an explainable AI to spot potential AI errors more easily. Conversely, when the loss is low, one prefers to use an unexplainable AI, which cannot be corrected, but has a higher baseline performance and does not necessarily need to be monitored.

The main goal of the paper is to establish what kind of contract a principal would offer to the agent given the different asymmetries of information -the choice of AI technology by the agent and the state of the world, the agent's correction effort. I suppose that the principal offers monetary transfers to the agent

and penalties if the agent provides erroneous recommendations. However, when the principal offers sufficient transfers to cover the cost associated with the correction of an explainable AI, the agent can extract a rent. Indeed, the agent knows when he can avoid verifying the AI's results, and exerts no efforts, but receives a high level of transfer. On the contrary, offering lower levels of transfers implies that the agent will not perform verification processes in the riskiest cases, as these processes will be costly as well. In that case, the level of rent extracted by the agent will be lower, but the social welfare, compounded by the sum of the agent's surplus and the principal profit falls as well. As a result, a distortion may appear, depending on the characteristics of the AI. Indeed, the principal can select a contract maximizing his own payoff, but that fails to reach the first best. Thus, asymmetries of information induce that the agent will not encouraged to use the correct type of AI, and the correct effort in terms of correction.

This paper is linked to a burgeoning literature that studies the use of AI from a principal-agent perspective. For instance, (Agrawal, Gans, and Goldfarb, 2018) or (Agrawal, Gans, and Goldfarb, 2019) study the potential complementary between the AI and the humans. Furthermore, Athey, Bryan, and Gans (2020), baed on Aghion and Tirole (1997), study the allocation of authority between an agent and a principal when the two parties have opposite interests McLaughlin and Spiess (2022), or Dell'Acqua (2021) study some agency problems regarding AI monitoring by agents, arguing that agents can become "lazy" when they can have access to an AI. However, contracts involving monetary transfer have not been evoked in this literature, which is the key point in this paper.

The rest of the paper is organized as follows. Section 2 is dedicated to the literature review. Section 3 introduces the model, Section 4 presents the characterization of the first best, and Section 5 the private equilibrium. In Section 6, I present some extensions of the baseline model. Last, in section 7, I conclude.

2 Related Literature

AI has a great impact on activities and will affect tasks requiring a high level of ability (Acemoglu and Restrepo, 2019; Brynjolfsson and McAfee, 2015; Autor, 2015). AIs are currently used in numerous areas, such as predictive justice (Sloan, Naufal, and Caspers, 2018; Ludwig and Mullainathan, 2021), finance (Abis and Veldkamp, 2020), or hiring decisions (Cowgill, 2018; Hoffman, Kahn, and Li, 2018) have already been documented. The literature not only focuses on the technical aspects of AI but also on the manner human behavior regarding decision-making evolves. Indeed, regarding decision-making, there is a clear distinction to make between the *predictive algorithm* and the *decision rule* associated with this algorithm (Rambachan et al., 2020). AI provides recommendations for action. However, only humans can formalize the payoffs

associated with the subsequent decisions. As a consequence, scholars have highlighted some potential harms associated with AIs (Acemoglu, 2021). For instance, a large strand of the literature focuses on the question of algorithmic fairness (Gillis, McLaughlin, and Spiess, 2021; Gillis, 2020; Kleinberg et al., 2018b), or the limits and design of algorithmic audits on group-specific screening (Meursault et al., 2022). Indeed, algorithms can discriminate against sensible groups, such as black people or women. Indeed, algorithms are used to improve the overall accuracy of decisions and do not integrate the limitation of bias as an objective, resulting in a social loss.

As a consequence, one of the great challenges associated with the use of AI is to spot the AI errors and correct them (Cowgill and Tucker, 2019). However, reducing the occurrences and the magnitude of these losses needs some human intervention. Related to fairness concerns, an other strand of the literature studies the complementary between humans and algorithms in the context of decision-making (Donahue, Chouldechova, and Kenthapadi, 2022; Bansal et al., 2021). In particular, scholars have highlighted the issues linked to the explainability of algorithms (Beaudouin et al., 2020). Indeed, some of the promising algorithms are also "black boxes", providing results without explanations Kleinberg et al. (2018a). Doshi-Velez and Kim (2017), Lundberg and Lee (2017) or Hamon et al. (2022) argue that explainability may be necessary for ethical and safety purposes, and to fix mismatched objectives. Designing an explainable AI may also be mandatory to prevent serious errors (Lipton, 2017; Amodei et al., 2016; Kleinberg, Mullainathan, and Raghavan, 2016). However, designing an algorithm to introduce an explainability objective decreases its overall level of accuracy. Indeed, algorithmic accuracy and explainability are often presented as trade-offs: increased explainability leads to decreased accuracy and vice versa. However, most algorithms were built with only accuracy in mind. Even algorithms with explanation models built on top may lead to sacrifices in accuracy in some situations. Thus, a decision maker should select the "right" level of explanation returned by his model, balancing the accuracy objective and the explainability requirements (Beaudouin et al., 2020). However, the regulation associated with the accountability of AI that should be applied to AI holders is still under discussion (Busuioc, 2020). Furthermore, note that the accuracy of a system combining AI and humans is better when the AI can provide explanations to human teams (Lundberg et al., 2018; Lai and Tan, 2019; Schmidt and Biessmann, 2019). However, in Computer Science literature, the question of the costs associated with the correction of the AI is usually not considered as it would be in economics. Thus, Computer Science does usually not consider "economic" incentives for decision-makers to correct their algorithms.

Furthermore, scholars in economics try to design a general framework, aiming to analyze the use of AI within an organization.Agrawal, Gans, and Goldfarb (2018) or Agrawal, Gans, and Goldfarb (2019), study

the complementary and the substitutability between AI and humans. Regarding the use of AI in a principal-agent model Athey, Bryan, and Gans (2020), exploits a usual principal-agent model by Aghion and Tirole (1997) to study the allocation of authority between an agent and an AI when the principal has access to an AI that can provide free advice. Dell'Acqua (2021) shows that the use of an AI by a principal can incentivize the agent to exert less effort than necessary. Blattner, Nelson, and Spiess (2023) studies how a principal can control the type of AI used by an agent by asking for some explanation from him. Lastly, McLaughlin and Spiess (2022) studies the situation where an AI not only provides a recommendation of action but also alters the preferences of a decision maker; the conclusion is that it may be preferable not to disclose AI recommendation to encourage the agent to make its efforts.

However, the situation where the principal offers direct compensation mechanisms is not evoked in these papers. Thus, I study a model involving decision-making with moral hazard with direct transfers between the principal and the agent, with no private values of the project for the agent. Thus, this paper is linked to some usual models in the principal agent literature, as Holmstrom (1979) or Maskin and Tirole (1990). Regarding AI technology, I use the idea, already present in the principal-agent literature, that restricting agent choice while still leaving enough flexibility to leverage the agent's technology and private information can be useful and would increase the overall social welfare (Levitt and Snyder, 1997; Szalay, 2005; Alonso and Matouschek, 2008)

3 Model

3.1 An Example

In this section, I suppose that a principal (he) delegates a decision to an agent (she). Both are risk-neutral. The principal faces a decision, that leads to positive or negative payoffs for him. The agent uses an AI to provide the recommendation. The recommendation can be either good or bad, leading to positive or negative payoffs respectively. The possible payoffs depend on the state of the world, which is privately known by the agent. For instance, according to Agrawal, Gans, and Goldfarb (2018) a bad decision may engender a "hidden cost". When a hidden cost happens, the negative payoff will be higher than when there is no hidden cost. However, spotting potential hidden costs requires some expertise, implying that only the agent can determine the state of the world. The state of the world conditions the technology used by the agent.

To be clearer, let's take an example. AI is currently used to select portfolios. Uninformed consumers

ask a financial manager to design a portfolio. The AI-designed portfolio can be well-performing (i.e. good decision) or low-performing (i.e. bad decision). However, the eventual returns of the portfolio depend on the AI's recommendation but also on the state of the economy, which is unknown by the consumer, but known by the manager. Indeed, bad investments have anecdotal consequences when the economy goes well but can have catastrophic ones when the economy experiences a crisis. In the terminology of Agrawal Gans and Goldfarb, a crisis will be a "hidden" cost. The portfolio manager can monitor the portfolio created by the AI, or only follow the AI's recommendation. Monitoring means abandoning the AI portfolio and reverts to a secure portfolio, with lower returns and lower risks. The manager monitors the portfolio thanks to his expertise. However, I suppose that the manager has to understand the reasons why the AI selects a given decision when accepting or rejecting the AI's recommendation.

An illustration of this can be found in Blattner, Nelson, and Spiess (2023). They argue that restricting the complexity of algorithms and relying on simpler, fully transparent decision rules. Indeed, understanding AI's models, and revealing the key variables allows us to reveal when the AI makes a costly error. If the manager sees that the AI selects a decision that is not adapted to the situation (for instance, if the AI gives important weight to the expected return of an asset, and neglects its level of risk), the manager can spot it and go back to a more secure area.

3.2 Model

I suppose that the agent and the principal both have linear functions of utility. There are two states of the world: a risky state and a safe state. The risky state happens with a probability q. The loss associated with a bad decision is equal to V. The safe state happens with a probability $1 - q$. The loss associated with a bad decision is equal to 0. In both cases, a good decision leads to a payoff equal to 1. The table of payoffs is as follows:

	1-q	**q**
Good Decision	1	1
Bad Decision	0	-V

Once the agent knows the state of the world, he selects an AI to obtain a recommendation. I suppose that the agent can select an AI among different types, without enabling or training costs. In the model, two types of AI are available, an explainable AI and an unexplainable one.

The explainable AI, recommends a good decision with a probability p, and a bad one with a probability $1-p$ with $p > \frac{1}{2}$. If the agent uses an explainable AI, he can exert a costly effort *ex-post* to correct the AI's recommendations in order to tackle costly errors. Eventually, the agent can reject the AI's recommended project if it appears that the AI's recommendation is erroneous. In that case, the principal's payoff is equal to 0 as no decision has been taken. The probability of spotting an AI error is equal to e. I suppose that the cost associated with an effort e is convex and is equal to αe^2.

If the risky state arises, the payoff associated with a bad decision is equal to $-V$. Thus, for a correction effort e, the payoff associated with the explainable AI is:

$$p - V(1-p)(1-e) - \alpha e^2 \tag{1}$$

Note that e is between 0 and 1. If the agent's analysis concludes that a recommendation is good, the agent sends it to the principal, leading to a payoff equal to 1. Otherwise, if the agent finds out that the AI's recommendation is erroneous, he does not send any recommendation, and the payoff of the principal will be equal to 0.

Conversely, if the safe state arises (i.e. the payoff associated with a bad decision is equal to 0), the overall payoff associated with the explainable AI is used, depending on e is:

$$p - \alpha e^2 \tag{2}$$

Indeed, in a safe state, making a bad decision implies no losses for the principal. Thus, there is no interest in providing any effort in that state.

The agent can also use an unexplainable AI to perform a recommendation. I assume that the unexplainable AI delivers a more accurate prediction than the explainable one. Indeed, as explained above, unexplainable AIs usually offer higher performances compared with more explainable algorithms. In the model, it will be translated by the assumption that unexplainable AI delivers a good recommendation more frequently than explainable ones. Thus, the probability that an unexplainable AI delivers a good recommendation equal to $p + k$, with $k > 0$. Conversely, the probability that the unexplainable AI delivers an erroneous recommendation is equal to $1 - p - k$.

If the payoff associated with a bad decision is equal to 0, the expected payoff associated with the use of the unexplainable AI is equal to $p + k$. Conversely, if the payoff associated with a bad decision is $-V$, the expected payoff associated with the unexplainable AI is:

$$p + k - V(1 - p - k) \tag{3}$$

3.3 Different Type of Contracts

3.3.1 Transfer Contract

I suppose that the principal can neither use an AI himself nor determine the state of the world. He must rely on an agent to make a decision. The principal is informed about the different types of AI and their accuracy. However, the type of AI that is actually used is privately known by the agent. I assume that the principal can offer 3 different types of contracts to the agent:
The first type of contract is defined as follows: The principal offers a positive transfer t whenever the agent delivers a recommendation. However, in exchange for such a transfer, the agent must pay a penalty if the principal incurs a loss. I suppose that the agent does not face any budget constraints. Thus, the value of the penalty can be freely determined by the principal, as long as it does not exceed the value of the loss V.

To ensure the agent's participation, the principal must offer a sufficient transfer to encourage the agent to provide a recommendation in the risky state. The transfer has to be high enough to cover the costs associated with the effort exerted by the agent, and the expected value of the penalty that can be paid by the agent. In the safe state (i.e. when the possible loss is equal to 0), offering this type of contract incentivizes the agent to use the unexplainable AI. Indeed, the transfer t occurs, and the agent does not face any costs when he uses the unexplainable AI. Conversely, in the risky state, the possibility of paying a penalty may incentivize the agent to use an expainable AI and discard the AI's errors.

3.3.2 Unexplainable AI contract

The second type of contract (called unexplainable AI contract) is as follows: The principal can offer no monetary transfers, but no penalty. Indeed, as I assumed that unexplainable AI is costless. Thus, there is a second type of contract, where the principal asks for the unexplainable AI's recommendation for the two states. The principal does not have to offer any transfer as the agent can provide a recommendation from the unexplainable AIs without any costs. Unlike the first type of contract, the agent does not extract any rent.

3.3.3 Explainable AI contract

A third type of contract, called explainable AI contract exists. The principal can control the effort exerted by the agent. Here, I suppose that the principal can perfectly verify the effort exerted by the agent. Some mechanisms can be imagined. For instance, the principal can ask for explanations in addition to the recommendation, as postulated in Blattner, Nelson, and Spiess (2023) or Caro and Nelson (2024) for instance. However, this type of contract enforces the use of the explainable AI, as the agent cannot justify any efforts if he selects the unexplainable one.

Thus, with this third type of contract, the principal imposes a given level of effort $\tilde{e}$ to the agent. In that case, the agent always selects the explainable AI, regardless of the state of the world, and exerts an effort $\tilde{e}$. In exchange for this effort, the principal repays the cost associated with the effort, namely $\alpha(\tilde{e})^2$. As in the second type of contract, the agent does not extract any rent from the principal.

4 Characterization of the First Best

In this section, I determine which type of AI should be used, depending on the state of the world. Moreover, I determine the optimal effort when an explainable AI is eventually used.

When the payoff associated with a bad decision is $-V$, the use of the explainable AI is valuable, as it reduces the probability of obtaining a negative payoff. In proposition 1, I determine the optimal effort that should be exerted when the possible loss is equal to $-V$.

Proposition 1. *In the risky state, the optimal effort when the explainable AI is used is* $e^* \equiv \frac{V(1-p)}{2\alpha}$ *if* $\frac{V(1-p)}{2\alpha} < 1$*, and* $e^* \equiv 1$ *otherwise.*

Proof. In the risky state, the loss is equal to $-V$, the expected payoff associated with the use of the explainable AI is:

$$W_V = p - V(1-p)(1-e) - \alpha e^2$$

Taking the derivative with respect to e gives:

$$\frac{dW_V}{de} = V(1-p) - 2\alpha e$$

Solving for the first order condition of pay-off maximization gives:

$$e^* \equiv \frac{V(1-p)}{2\alpha}$$

If the value of $\frac{V(1-p)}{2\alpha}$, is higher than 1, the constraint is satiated. In that case, the optimal level of e is equal to 1. It is optimal to avoid all risks of losses. In that case, the welfare is equal to $p - \alpha$. □

Exerting a correction effort is costly; as the baseline accuracy of the unexplainable AI is higher than the explainable AI's one, the use of the unexplainable AI is valuable if and only if V is high enough. We determine the conditions such that the explainable AI is preferred to the unexplainable AI in the risky state.

Corollary 1. *If $V \geq \frac{V(1-p)}{2\alpha}$, the use of the explainable AI is preferred to the use of the unexplainable AI if and only if the value of k is lower than $\frac{(1-p)V-\alpha}{V+1}$.*

If $V < \frac{V(1-p)}{2\alpha}$, the use of the explainable AI is preferred to the use of unexplainable AI if the value of k is lower than $\frac{(1-p)V-\alpha}{V+1}$.

Thus, if V is high enough, the first best is to use the explainable AI in the risky state and the unexplainable one otherwise.

For the next section, I make the two following assumptions:

Assumption 1. *The value of V is higher than $\frac{2\alpha}{1-p}$*

Assumption 2. *The value of k is lower than $\frac{(1-p)V-\alpha}{V+1}$.*

According to Proposition 1 and Corollary 2, these assumptions have the following implications on the first best:

1. If the loss associated with a bad decision is equal to $-V$, the optimal effort e is equal to 1.
2. If the loss associated with a bad decision is equal to $-V$, it is better to use the explainable AI.

Note that the first assumption is not crucial for the future statement of the results. I will relax this assumption in the extensions. The second one cannot be relaxed. Indeed, if the assumption does not hold, it implies that the first best is using the unexplainable in every situation.

I determine which type of contract can lead to the first best. As specified in the section above, offering a transfer and a penalty implies that the agent uses an explainable AI and performs a correction effort in the risky state. Conversely, in the safe state, the agent uses the unexplainable AI. Thus, if the couple transfer/penalty proposed by the agent ensures that the agent performs the optimal effort $e^* \equiv 1$, the offered contract would lead to the first best.

Proposition 2. *Under Assumption 1 and Assumption 2, the first best is reached if the transfer is $t \equiv \alpha$ and the penalty set by the principal is equal to V*

Proof. It is straightforward that if such a contract is proposed, the agent uses the unexplainable AI in the safe state and the explainable AI in the risky one. In the risky state, the payoff of the agent is:

$$\pi_A^R = t - V(1-p)(1-e) - \alpha e^2$$

Maximizing this agent's profit implies an effort $e = 1$, under Assumption 1. Thus, the effort chosen by the agent is $e \equiv 1$, which is the first best. □

Thus, the principal can reach the first best by offering a transfer/penalty contract. This contract will be called the first best contract for the rest of the section.

However, as specified above, the agent extracts a rent if the first best contract is proposed. Thus, the principal may prefer to propose another type of contract, decreasing the social welfare, but leaving no rent to the agent. In the next section, I determine the conditions on the correction costs and the relative accuracy of the explainable and unexplainable AI such that the principal prefers to offer a first-best contract rather than one of the two other possible contracts.

5 The Private Equilibrium

Offering the first best contract implies letting a rent to the agent. As the transfer is equal to α, the agent always exerts an effort equal to 1 in the risky state. As the transfer offered by the principal fully covers the cost associated with the effort. Thus, if the loss is equal to V, the agent extracts no rent and provides a recommendation, leading to a positive payoff for the principal. However, in the safe state, the agent still receives a transfer α but exerts no effort as he will use the unexplainable AI to provide a recommendation. As there are no losses associated with an error, the agent always extracts a rent α.

We need to compare the different payoffs that can be obtained by the principal. Recall that the accuracy of the unexplainable AI is $p+k$, and the accuracy of the explainable AI is p. First, I derive the condition on k upon which the principal prefers to offer the first best contract rather than asking for the unexplainable AI recommendation only. Indeed, the higher the value of k, the lower the loss endured by the principal when the cost associated with an error is equal to $-V$. Conversely, as both contracts imply that the agent uses the unexplainable AI in the safe state, an increase or a decrease in k does not modify the difference between

the principal's profit obtained with the two contracts.

For the principal, an increase of k increases the profit associated with the unexplainable AI's contract relative to the profit associated with the first best contract. As a consequence, if k is high enough, the principal prefers to ask for the unexplainable AI's recommendation only. Indeed, the disutility faced by the principal when he receives the payoff -V will be lower than the level of rent left to the agent.

Proposition 3. *The principal prefers to offer the first best contract rather than asking for the unexplainable AI recommendation if and only if* $k < \bar{k} \equiv \frac{Vq(1-p)-\alpha}{qV+q}$. *Otherwise, the principal always asks for the unexplainable AI's recommendation.*

Proof. When the principal offers a transfer $t \equiv a$ to the agent, the principal's profit is:

$$\pi_P^{FB} \equiv p + (1-q)k - a \tag{4}$$

When the principal only asks for unexplainable AI's recommendations, the principal profit is:

$$\pi_P^{U} \equiv p + k - qV(1-p-k) \tag{5}$$

Thus, we have $\pi_P^{FB} = \pi_P^{U} \longleftrightarrow k < \bar{k} = \frac{Vq(1-p)-\alpha}{qV+q}$ □

When the accuracy of the unexplainable AI is much higher than the accuracy of the explainable AI, the first best contract will not be selected by the principal. It implies that the first best is not reached, leading to a distortion.

The profile of the payoffs is given in the graph below. The dashed red line represents the social welfare, which is equal to here, to the first best).

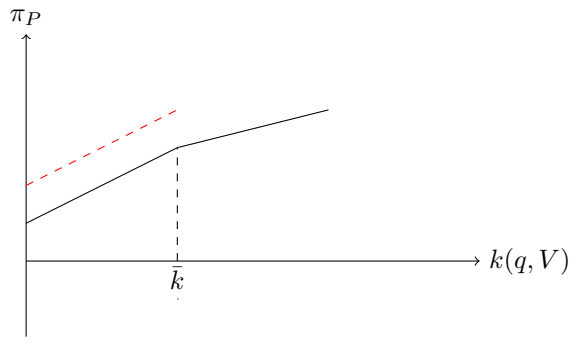

Figure 1: $q > \frac{2\alpha}{V(1-p)}$

Now, I derive the conditions upon which the principal prefers to offer the first best contract rather than imposing an effort $\tilde{e}$ to the agent in all states. Indeed, when the principal imposes a given level of effort $\tilde{e}$ the level of effort is lower than the optimal level of effort when the loss is equal to V and of course higher than the optimal level of effort when the loss is equal to 0. Thus, imposing $\tilde{e}$ implies a social loss, due to the suboptimal effort exerted in the two states. However, this kind of contract implies that no rent is left to the agent. Thus, if the overall loss for the principal is lower than the transfer left to the agent, the principal prefers to impose an effort on the agent.

Furthermore, an increase in k increases the principal's payoff associated with the first best contract, but not the principal's payoff if he proposes the third type of contract. Thus, there is naturally a minimum value of k such that the principal prefers to offer the first best contract.

In Proposition 4, I show that if the value of V is too important, the principal always prefers to offer the first best contract.

Proposition 4. *If $q \geq \frac{2\alpha}{V(1-p)}$, the principal prefers to offer the first best contract for all value of k*

Proof. The optimal level of $\tilde{e}$ chosen y the principal should maximize the following expression:

$$\pi_P^E \equiv p - a\tilde{e} - (1-p)qV(1-\tilde{e}) \tag{6}$$

The first order condition gives:

$$\frac{d\pi_P^E}{d\tilde{e}} = qV(1-p) - 2\alpha\tilde{e} \tag{7}$$

Solving for the first order condition gives $\tilde{e}^* = \frac{Vq(1-p)}{2\alpha}$. If $q \geq \frac{2\alpha}{V(1-p)}$, the optimal value of $\tilde{e}$ should be higher than 1, which is impossible. In that case, $\tilde{e}^* = 1$.

In that case, it is always better for the principal to offer the first best contract. Indeed, in the risky state, the payoff associated with the two different contracts is the same, as the effort is equal to 1 in both cases. However, if the loss associated with a bad decision is equal to 0, the principal receives a payoff equal to $p-\alpha$, as the agent still uses an explainable AI because the principal enforces it. However, if the principal offers the first best contract, his payoff will be equal to $p + k - \alpha$, as the agent can use the unexplainable AI in this situation. □

If the value of V is too important, or the probability that the risky state arises is too high, the principal will set $\tilde{e} = 1$. The overall cost in the two states will be equal to α. Thus, the third type of contract is clearly

less profitable than the first best contract for the principal.

If the risky state arises less often, (or if the loss is quite low) the principal prefers to set an effort $\tilde{e} < 1$. Indeed, a decrease in $\tilde{e}$ increases the payoff of the principal when the payoff associated with an error is null, but decreases the payoff of the principal when the payoff associated with an error is negative. If q is sufficiently low, the first effect dominates the second, and the principal prefers to set a lower $\tilde{e}$. Thus, we will have $\tilde{e} < 1$. In other words, the principal experiences a loss $(1-p)(1-\tilde{e})$ with a probability q but will receive a higher payoff when the cost associated with an error is null. Note that, if the third type of contract is selected, the value of k does not impact the principal's payoff. However, it positively impacts the principal's payoff when he offers the first best contract. Thus, for k sufficiently high, the principal would prefer to offer this type of contract.

Proposition 5. *If $q < \frac{2\alpha}{V(1-p)}$, the principal prefers to offer the first best contract rather than imposing a given level of effort if and only if $k > \underline{k}$*

$$\underline{k} \equiv \frac{(2\alpha - (1-p)qV)^2}{4\alpha(1-q)} \tag{8}$$

Proof. As shown in the proposition above, the optimal effort when the
principal imposes that the agent exerts a given level of effort $e^* \equiv \frac{(1-p)Vq}{2\alpha} < 1$. Thus, we have:

$$\pi_P^E \equiv p - ae^{2*} - (1-p)qV(1-e^*) \tag{9}$$

We have $\pi_P^E = \pi_P^T \longleftrightarrow k \equiv \underline{k} = \frac{(2\alpha-(1-p)qV)^2}{4\alpha(1-q)}$ □

The principal prefers to offer the first best contract rather than imposing an effort $\tilde{e}$ if the value of k is sufficiently high. Indeed, the value of k positively impacts the profit associated with the first best contract, as it increases the payoff of the principal in the safe state. Moreover, according to Proposition 3, the principal offers the first best contract if and only if k is below a given threshold $\bar{k}$. It implies that we can deduce the values of k such that the principal selects the first est contract.

Proposition 6. *Thus, the principal offers the first best contract if and only if k is in the interval $[\underline{k}, \bar{k}]$. Otherwise, the principal selects a contract that leads to sub-optimal outcomes from a social welfare perspective.*

Proof. It directly comes from Proposition 3 and Proposition 5. □

The profiles of the different payoffs, depending on k are presented below.

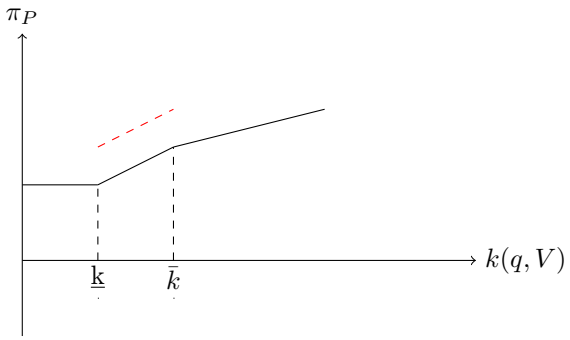

Figure 2: $q \in [\underline{q}, \frac{2\alpha}{V(1-p)}]$

5.1 Comparative Statics

Now, I study how $\bar{k}$ and $\underline{k}$ evolve with the probability that the risky state arises q. First, we assess how the value of $\bar{k}$ evolves relative to q

Lemma 1. *The value of $\bar{k}$ increases with q.*

Proof. We need to show that the derivative of $\bar{k}$ relative to q is positive. The derivative of $\bar{k}$ relative to q is:

$$\frac{d\bar{k}}{dq} = \frac{\alpha}{(V+1)q^2} > 0 \tag{10}$$

□

Indeed, the higher the probability that a loss can happen, the higher the profit associated with the first best contract relative to the unexplainable AI contract.

Now, we assess the effect of q on $\underline{k}$

Lemma 2. *The value of $\underline{k}$ decreases with q*

Proof. We must show that the derivative of $\underline{k}$ relative to q is negative. As stated in Proposition 5, $\underline{k} \equiv \frac{(2\alpha-(1-p)qV)^2}{4\alpha(1-q)}$. The derivative of $\underline{k}$ relative to q is:

$$\frac{d\underline{k}}{dq} = \frac{(2\alpha - (1-p)V(2-q)))(2\alpha - (1-p)Vq)}{4\alpha(1-q)^2} < 0 \tag{11}$$

Here, I supposed that $q < \frac{2\alpha}{V(1-p)}$, the second term of the numerator is positive. Regarding the first term of the numerator, we have $2\alpha - (1-p)V(2-q)) > 2\alpha - (1-p)V)$. Again, by assumption, this expression is lower than 0. Thus, the product is negative, which ends the proof.

□

Indeed, when the probability that the risky state happens increases, the payoff associated with the first best contract increases relative to the payoff associated with the explainable AI contract. Indeed, by selecting the first best contract, the principal is fully covered against all the potential losses. Thus, an increase in q does not impact negatively the principal's payoff on the losses side. If the principal selects the first best contract. However, an increase in q implies that the loss V will happen more frequently to the principal if he constrains the agent's effort to $\tilde{e} < 1$.

Thus, we can deduce the effect of q on the principal's choice of contracts.

Proposition 7. *There is a value $\underline{q}$ such that $\underline{k} = \bar{k}$. If $q < \underline{q}$, the principal imposes a level of effort $e \equiv \frac{q(1-p)V}{2\alpha}$ if $k \geq \tilde{k} \equiv \frac{(2-q)(qV^2)(1-p)^2}{4(\alpha qV+\alpha)}$.*
Otherwise, the principal asks for the unexplainable AI's recommendation.

Proof. As shown in the Lemma 5, $\underline{k}$ is increasing in q, and $\bar{k}$ is decreasing in q. Thus, there is a value $q \equiv \underline{q}$ such that $\bar{k} = \bar{k}$.

If $q < \underline{q}$, offering a transfer against a penalty is never an option for the principal, for all values of k. Indeed, if $k > \bar{k}$, the principal prefers to offer a transfer rather than ask for the unexplainable recommendation only. However, as $bar k < \underline{k}$, the principal prefers to impose a given level of effort rather than offer a transfer. Conversely if $k < \bar{k}$, the principal prefers to offer a transfer rather than impose a given level of effort rather than offer a transfer. However, as $bar k < \underline{k}$, the principal prefers to ask for the unexplainable recommendation only rather than offering a transfer. Thus, the principal only considers the options of imposing an effort $\tilde{e} \equiv \frac{V(1-p)q}{2\alpha}$ to the agent, or asking for the unexplainable AI's recommendation. The principal compares the two profit π_P^E and π_P^U, and we have:

$$\pi_P^E > \pi_P^U \longleftrightarrow k < \tilde{k} \equiv \frac{(2-q)(qV^2)(1-p)^2}{4(\alpha qV+\alpha)} \tag{12}$$

□

There is a value of q such that the first best contract is never preferred by the principal, regardless of the value of k. In that case, the principal asks for unexplainable AI if k is sufficiently high, and imposes an effort $\tilde{e}$ otherwise. The subsequent profile of profit is given below:

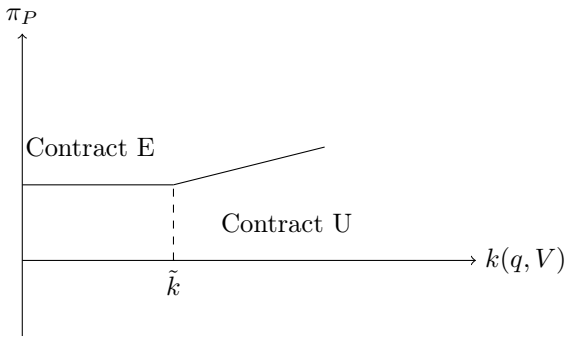

Figure 3: $q < \tilde{q}$

6 Continuous Loss

Now, I suppose that there is a continuum of loss. The principal faces a project. Taking a good decision leaves a payoff of 1, and making a bad decision implies a loss of V, with V following a distribution function $f(v)$ over a support range of $[0, \overline{V}]$. This new specification of the losses adds more complexity to the model-and subsequently modifies how the contracts are settled between the agent and the principal.

The value of V is privately known by the agent. The principal can offer a penalty, whose value depends on the realized value of the loss V. As in the previous section, the value of the penalty is bounded by V. The principal can tailor a "menu" of penalties, depending on the loss he receives when the project fails. If we refer to the previous section, the way to achieve the first best could be to offer $t = \alpha$ and a penalty that equals the value of the actual loss. Indeed, with this specification, the couple penalty-losses (i.e. the first best contract in the previous section) chosen by the principal does not necessarily lead to the social optima. Indeed, in this setting, the principal might not propose a transfer equal to α, and a penalty equal to the actual value of the loss, implying that the agent always delivers a recommendation and exerts the optimal level of effort when he uses the explainable AI. The principal can propose a remuneration lower than α but also a lower penalty value. Note that the optimal level of effort is exerted by the agent if and only if the level of penalty is equal to the eventual level of loss. Thus, the principal diminishes the rent but implicitly decreases the efforts of the agent in some states. Thus, the contract Moreover, the principal still compares the payoff associated with the "second best" contract with the type-one and type-two contracts.

6.1 Characterization of the First Best

First, I determine the threshold value $\tilde{V}^*$ such that when $V > \tilde{V}^*$, it is socially preferable to use the explainable AI. Otherwise, if $V \leq \tilde{V}^*$, it is socially optimal to use the unexplainable AI. Moreover, if

$V > \tilde{V}^*$, the agent should be incentivized to provide the optimal level of effort e^* (e.g., the level of effort that was defined in Proposition 1).

If we follow the logic that we used in the previous section, the best contract, from a social welfare perspective involves transfer and penalty. To encourage the agent to provide an effort in all situations, the transfer should be equal to α, to cover the agent's costs when the effort is equal to 1. Moreover, the penalty should be equal to V, to encourage the agent to exert the optimal effort.

In the next proposition, I prove that it is impossible to ensure that the agent uses the explainable AI if and only if $V > \tilde{V}^*$ and the unexplainable AI otherwise and that the agent exerts the optimal effort e^* when he uses the explainable AI. It implies that the first best cannot be reached with a contract involving monetary transfers.

Proposition 8. *Suppose that $V = \tilde{V}^*$. The is a level of penalty β such that the agent is indifferent between using the explainable AI and using the unexplainable AI is lower than* 1*.*

Proof. The transfer is equal to α regardless of the technology used by the agent.
Suppose that the penalty is equal to the level of losses for all V. In that case, the effort of the agent is equal to the first best effort $e^* \equiv \frac{V(1-p)}{2\alpha}$. The payoff of the agent when he uses the explainable AI is:

$$\alpha - (1-p)(1-e^*) - \alpha e^{*2} \tag{13}$$

This profit should be compared with the agent's payoff when the unexplainable AI is used (e.g. $\alpha - (1-p-k)V$). We have:

$$\alpha - (1-p)(1-e^*) - \alpha e^{*2} > \alpha - (1-p-k)V \Leftrightarrow V > \hat{V} \equiv \frac{4\alpha k}{(1-p)^2} \tag{14}$$

If $V < \hat{V}$, when the value of the penalty is equal to the value of the loss, the agent prefers to use an unexplainable AI. Otherwise, the agent uses the explainable AI, and the effort is equal to the optimal effort e^*.
However, we proved that the value of $\tilde{V}^*$ is

$$\frac{2\alpha k}{(1-p)^2} + \frac{2\sqrt{(ak)^2 + (1-p)^2)}}{(1-p)^2}$$

It is straightforward that $\hat{V} < \tilde{V}^*$. T Thus, if V is between $\tilde{V}^*$ and $\hat{V}$, the agent prefers to use the explainable

AI and set an effort e^* whereas it is socially optimal to use the unexplainable AI. □

The first best cannot be reached. Indeed, if $V \in [\tilde{V}, \tilde{V}^*]$ the principal prefers use the explainable AI and

This issue cannot be helped. Indeed, setting a penalty lower than the V experienced by the principal implies a dead weight loss for the principal. Thus, he is never incentivized to do it, and selects a penalty value equal to V.

6.2 The Private Equilibrium

Suppose that the principal offers the following contract: he sets $t = \alpha$ and a penalty equal to the value of the loss. In that case, the agent always provides a recommendation and always performs the optimal effort when the explainable AI is used. The profit obtained with $t = \alpha$ and full liability is:

$$\int_0^{\hat{V}} (p + k - \alpha) f(v) + \int_{\hat{V}}^{\bar{V}} (p - \alpha) f(v) \tag{15}$$

Due to the nature of the bounds, this is not the first best contract. However, this contract approximates the first best. Can the principal increase its profit by selecting another strategy? Indeed, when looking at the strategy presented above, it is clear the agent extracts a rent. The principal can decrease the level of rent by decreasing the level of transfer t. Thus, the principal can offer $t < \alpha$. However, by doing so, the principal discourages the agent from providing a recommendation in the states where V is high. Thus, if the principal proposes $t < \alpha$, he may also decrease the level of the penalty inflicted on the agent, to ensure her participation.

Lemma 3. *For a given proposed transfer $t < \alpha$, the agent accepts to send a recommendation if and only if the value of the penalty is lower or equal than β, with:*

$$\beta(t) = \frac{2(\alpha - \sqrt{a^2 - at})}{1 - p} \tag{16}$$

Proof. The value of β is the value that equals the profit of the agent to 0. As, for a given β, the effort of the agent is equal to $e^* = \frac{(1-p)\beta}{2\alpha}$ namely:

$$\pi_A(\beta, t) = t - \beta(1 - p)(1 - e^*) - \alpha e^{*2} \tag{17}$$

This value is equal to 0 for $\beta(t) = \frac{2(\alpha - \sqrt{a^2 - at})}{1-p}$ □

The value of β cannot be higher than V. If the value of β found above is above V, the constraint is satiated, and $\beta(t, V)$ is equal to V. In that case, the agent extracts a rent. Indeed, her payoff will be strictly higher than 0. However, the principal is fully covered against the losses. If the value of β found above is lower than V, it implies that the principal can constrain the agent's profit to 0. However, as long as $\beta < V$, the agent will perform a suboptimal level of effort. Indeed, only a penalty equal to the actual level of loss ensures that the optimal level of effort is exerted by the agent.

Thus, we can establish the profit of the principal, for a given transfer t and his associated profile of penalties $\beta(V)$.

1. If the value of the loss is under a given threshold $\tilde{V_1}$, the agent prefers to use the unexplainable AI
2. If the value of the loss is between $\tilde{V_1}$ ad $\tilde{V_2}$, the principal sets a penalty $\beta = 1$. The effort of the agent is the optimal level of effort, and the agent extracts a positive level of surplus
3. If $V > \tilde{V_2}$, the principal sets a penalty equal to $\beta < V$,. The value of the penalty imposes that the profit of the agent is equal to 0. The principal stops when the expected payoff is equal to 0.

The value $\tilde{V_1}$ is constant and equal to

$$\tilde{V_1} = \frac{2\alpha k}{(1-p)^2} \tag{18}$$

The value $\tilde{V_2}$ is attained when the payoff of the agent is equal to 0, when β is equal to 1 ad the agent uses the explainable AI.

$$\tilde{V_2} = \frac{2(\alpha - \sqrt{a^2 - at})}{1-p} \tag{19}$$

If $V > \tilde{V_2}$, the value of β such that, for t, the agent's profit is equal to 0 is:

$$\beta(t) = \frac{2(\alpha - \sqrt{a^2 - at})}{1-p} < V \tag{20}$$

The integral representing the principal's profit will be split in three. If V is low, the agent prefers to use the unexplainable AI the principal must determine a penalty β, with β lower than V to ensure the participation. In that case, the following conditions on β must hold.
The first condition implies that the agent's payoff, given t, is equal to 0.

The second condition guarantees that the principal's payoff, given β, V, and t, must be lower than 0.
If V is high enough such that there is no β such that the profit of the agent and the profit of the principal are positive, the principal prefers to set $\beta = 1$ for $V > \tilde{V}_2$. Thus, the agent does not send any recommendations for the highest values of V.

For each transfer that can be proposed by the principal, there is an optimal penalty vector B that maximizes the principal's profit

$$\pi_P(t) = \underbrace{\int_0^{\tilde{V}_1} (p+k-t)f(v)dv}_{\text{Unexplainable AI/Full Liability}} + \underbrace{\int_{\tilde{V}_1}^{\tilde{V}_2(t)} (p-t)f(v)dv}_{\text{Explainable AI/Full Liability}} + \underbrace{\int_{\tilde{V}_2(t)}^{\bar{V}(t)} (p-t-(1-p)(1-e^*(\beta(t)))(v-\beta(t)))f(v)dv}_{\text{Explainable AI/Limited Liability}} \tag{21}$$

$$\text{With : } \begin{cases} \beta(t) = \frac{2(\alpha-\sqrt{a^2-at})}{1-p} \\ e^*(\beta) = \beta\frac{1-p}{2\alpha} \end{cases}$$

The profit of the principal can be divided into three parts. The first part of the equation

The two first parts of the principal's profit are decreasing in t. Indeed, full liability is applied. Thus, the principal is covered against all losses.
The last part of the profit represents the situation where the principal cannot enforce full liability, as the agent's profit would fall below 0 if so. Thus, the principal must offer a penalty equal to β, lower than V. By increasing t, the principal can also increase the level of β in the last part of the profit, implying an increase in this section.

Thus, decreasing the value of t implies an increase in the two first part of the principal's profit and a decrease in the last part.

Proposition 9. *Depending on the distribution function $f(v)$, there is a value of $t < \alpha$ that locally maximizes the principal's profit. I call this profit $\pi_P^{t<\alpha}$. If the mass associated with the highest values of is sufficiently high, the principal sets $t = \alpha$ and $\beta = V$ for all V. Otherwise, the profit of the principal is $\pi_P^{t<\alpha}$.*

Sketch of the proof:

As seen in the lines above, the profit of the principal is compounded of three parts. The two first are decreasing in t, the second one is increasing in t. More precisely, the derivative of the two first terms relative

to t gives two linear in t. Conversely, given the values of e^* and β, taking the derivative of the third part regarding to t gives a convex term. Indeed, the expression of this part of the principal profit is a polynomial expression, with an order higher than 2. Thus, the derivative of the principal's profit is compounded by a negative linear term, and a positive convex term. Thus, if the distribution function fits well, there is a value of t such that the two terms equal each other. In particular, if the probability of reaching high losses is

If such a maximum exist, we must compare it with the principal profit when $t = \alpha$. Selecting a transfer slightly lower than α implies that the agent will never perform any recommendation is the expected loss is too high. Indeed, he will not perform the effort $e^* = 1$ for the high values of V. Moreover she will also will not provide any recommendation for the highest values of V. Thus, depending on the shape of $f(v)$ there is a discontinuity n the principal's profit around $t = \alpha$. As a consequence we must compare the values of $\pi_P^{t<\alpha}$ and the principal profit when $t = \alpha$.

In this proposition, I prove that the contract designed by the principal creates two distortions. First, selecting a transfer lower than α implies that the agent prefers not to provide a recommendation in the riskiest states. Moreover, if the transfer is lower than α, the agent will not perform the optimal effort for the highest values of V leading to a social loss. Thus, a double distortion may be induced by the principal-agent scheme.

7 Conclusion

In this paper, I have shown the different contracts that can be settled between an agent and a principal, given that different AI technologies can be used by the agent. There are asymmetries of information regarding AI technologies that can be used and the effort exerted by the agent. I show that these asymmetries create some distortions; indeed, due to the costless use of some AI technology, the agent can extract a rent. Indeed, the principal has incentives to push down the transfer provided to the agent to diminish the rent left to her. By doing so, the principal The result is fostered by the fact that unexplainable AIs are the most accurate ones. A quite well performing unexplainable AI technology incentivizes the principal to offer a contract leading to the use of the unexplainable AI only. Indeed, it diminishes the level of rent extracted but does not imply a great disutility due to loss consecutive to AI's errors for the principal. In this paper, I did not question some interesting points. First, the costs associated with the creation of the different need to be studied. Indeed, the results can change if we consider that different types of AIs are associated with such costs. Furthermore, I only considered the **private equilibrium** between the principal and the agent. Due to the risks associated with AIs, the intervention of a regulator to enforce a better use of AIs might be considered.

References

Abis, Simona and Laura Veldkamp. 2020. "The changing economics of knowledge production." *Available at SSRN 3570130* .

Acemoglu, Daron. 2021. "Harms of AI." Tech. rep., National Bureau of Economic Research.

Acemoglu, Daron and Pascual Restrepo. 2019. "Automation and new tasks: How technology displaces and reinstates labor." *Journal of Economic Perspectives* 33 (2):3–30.

Aghion, Philippe and Jean Tirole. 1994. "The Management of Innovation." *Quarterly Journal of Economics* 109:1185–1209.

———. 1997. "Formal and real authority in organizations." *Journal of political economy* 105 (1):1–29.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. "Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence." In *The economics of artificial intelligence: An agenda*. University of Chicago Press, 89–110.

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.

Alonso, Ricardo and Niko Matouschek. 2008. "Optimal Delegation." *Review of Economic Studies* 75 (1):259–293.

Amodei, Dario, Christopher Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. 2016. "Concrete Problems in AI Safety." *ArXiv* abs/1606.06565.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. "Machine bias." In *Ethics of data and analytics*. Auerbach Publications, 254–264.

Athey, Susan C, Kevin A Bryan, and Joshua S Gans. 2020. "The allocation of decision authority to human and artificial intelligence." In *AEA Papers and Proceedings*, vol. 110. 80–84.

Autor, David H. 2015. "Why are there still so many jobs? The history and future of workplace automation." *Journal of economic perspectives* 29 (3):3–30.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance."

Beaudouin, Valérie, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. 2020. "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach."

Blattner, Laura, Scott Nelson, and Jann Spiess. 2023. "Unpacking the Black Box: Regulating Algorithmic Decisions."

Boyer, Marcel and Jean-Jacques Laffont. 1997. "Environmental risks and bank liability." *European Economic Review* 41 (8):1427–1459.

Brynjolfsson, Erik and Andrew McAfee. 2015. "Will humans go the way of horses." *Foreign Aff.* 94:8.

Busuioc, Madalina. 2020. "Accountable Artificial Intelligence: Holding Algorithms to Account." *Public Administration Review* 81.

Caro, Spencer and Scott Nelson. 2024. "The Arity of Disparity: Updating Disparate Impact for Modern Fair Lending." *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2024-18).

Cowgill, Bo. 2018. "Bias and productivity in humans and algorithms: Theory and evidence from resume screening." *Columbia Business School, Columbia University* 29.

Cowgill, Bo and Catherine Tucker. 2019. "Economics, Fairness and Algorithmic Bias." *SSRN Electronic Journal* .

Davenport, Thomas and Jeanne Harris. 2017. *Competing on analytics: Updated, with a new introduction: The new science of winning*. Harvard Business Press.

Dell'Acqua, Fabrizio. 2021. "Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters." .

Donahue, Kate, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. "Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness."

Doshi-Velez, Finale and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." *arXiv: Machine Learning* .

Gillis, Talia, Bryce McLaughlin, and Jann Spiess. 2021. "On the fairness of machine-assisted human decisions." *arXiv preprint arXiv:2110.15310* .

Gillis, Talia B. 2020. "False Dreams of Algorithmic Fairness: The Case of Credit Pricing." URL `https://api.semanticscholar.org/CorpusID:219132610`.

Hamon, Ronan, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. 2022. "Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making." *IEEE Computational Intelligence Magazine* 17 (1):72–85.

Hoffman, Mitchell, Lisa Kahn, and Danielle Li. 2018. "Discretion in Hiring." *The Quarterly Journal of Economics* 133 (2):765–800.

Holmstrom, Bengt R. 1979. "Moral Hazard and Observability." *The Bell Journal of Economics* 10:74–91.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018a. "Algorithmic Fairness." *AEA Papers and Proceedings* 108:22–27.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores."

Kleinberg, Jon M., Jens Ludwig, Sendhil Mullainathan, and Cass Robert Sunstein. 2018b. "Discrimination in the Age of Algorithms." *DecisionSciRN: Non-Rational Decision-Making (Topic)* URL `https://api.semanticscholar.org/CorpusID:60440828`.

Lai, Vivian and Chenhao Tan. 2019. "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM.

Levitt, Steven D. and Christopher M. Snyder. 1997. "Is No. News Bad News? Information Transmission and the Role of "Early Warning" in the Principal-Agent Model." *RAND Journal of Economics* 28 (4):641–661.

Lipton, Zachary C. 2017. "The Mythos of Model Interpretability."

Ludwig, Jens and Sendhil Mullainathan. 2021. "Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System." *Journal of Economic Perspectives* 35 (4):71–96. URL `https://www.aeaweb.org/articles?id=10.1257/jep.35.4.71`.

Lundberg, Scott, Bala Nair, Monica Vavilala, Mayumi Horibe, Michael Eisses, Trevor Adams, David Liston, Daniel Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. 2018. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery." *Nature Biomedical Engineering* 2.

Lundberg, Scott M. and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Neural Information Processing Systems*.

Maskin, Eric and Jean Tirole. 1990. "The principal-agent relationship with an informed principal: The case of private values." *Econometrica: Journal of the Econometric Society* :379–409.

McLaughlin, Bryce and Jann Spiess. 2022. "Algorithmic Assistance with Recommendation-Dependent Preferences." *arXiv preprint arXiv:2208.07626* .

Meursault, Vitaly, Daniel Moulton, Larry Santucci, and Nathan Schor. 2022. "One Threshold Doesn't Fit All: Tailoring Machine Learning Predictions of Consumer Default for Lower-Income Areas." Working papers, Federal Reserve Bank of Philadelphia.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. "An Economic Perspective on Algorithmic Fairness." *AEA Papers and Proceedings* 110:91–95.

Schmidt, Philipp and Felix Biessmann. 2019. "Quantifying Interpretability and Trust in Machine Learning Systems."

Sloan, Carly Will, George S Naufal, and Heather Caspers. 2018. "The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes." IZA Discussion Papers 11948, Institute of Labor Economics (IZA). URL `https://ideas.repec.org/p/iza/izadps/dp11948.html`.

Szalay, Dezsö. 2005. "The economics of clear advice and extreme options." *The Review of Economic Studies* 72 (4):1173–1198.

Washington, Anne L. 2018. "How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate." *Colo. Tech. LJ* 17:131.