The Adoption of Artificial Intelligence at the System Level

Ajay Agrawal, Joshua S. Gans, Avi C. Goldfarb

December 2020

Abstract

TBD

* Rotman School of Management, University of Toronto and NBER.

** All correspondence to joshua.gans@utoronto.ca.

1 Introduction

Over the past decade, artificial intelligence (AI) has emerged as a potential general purpose technology (Cockburn et al. (2019)). Spurred on by advances in machine learning, the cost of prediction across various domains has started to fall at an accelerating pace (Agrawal et al. (2018a)). This raises interesting questions of where AI will be adopted and also its potential disruptive impact on employment and businesses (Gans and Leigh (2019)).

To date, our conception of AI adoption has mainly operated at the unit of a task or decision (e.g., Frank et al. (2019); Acemoglu and Restrepo (2018)). To forecast the potential impact of AI on employment, for example, there have been numerous exercises designed to identify jobs at risk from AI, the tasks that comprise jobs that are at risk and the more general impact of automation on the workplace (Webb (2020); Brynjolfsson and Mitchell (2017); Brynjolfsson et al. (2018); Felten et al. (2018)). That said, some of questioned whether this task-level approach is suitable. Bresnahan (2020) argues that AI is an information technology and, traditionally, such technologies have required organisational redesign to be fully adopted. This is readily apparent in patterns of adoption of earlier generations of IT (Bresnahan and Greenstein (1996); Bresnahan et al. (2002); Aral et al. (2012); Dranove et al. (2014)).

Bresnahan (2020) emphasizes that AI will only be adopted where the consequences of making errors in decisions have low stakes and where the decisions themselves are contained, say, within specific modularised functions. He argues that this is the case for AI that primarily enhances the functionality of user interfaces and ones that sort information prior to it being acted upon by others. Thus, while Bresnahan's assessment of the consequences of AI are different from the task-based model, his predictions about which organizations will be early adopters are similar: modular organizations where the consequences of a mistaken machine prediction are relatively low.

As an example, Bresnahan discusses Amazon's recommendation engine. The recommendation engine uses data about consumer purchasing to predict which items a particular consumer is likely to purchase next. These items are shown to the consumer, who then chooses which item(s) to purchase (if any). This AI system is modular in the sense that it does not require substantive changes elsewhere in the organization. It is a change to the information each consumer sees, but the path from information to purchase to delivery is unchanged. Furthermore, this AI system is low stakes. Here, Bresnahan emphasizes the specific nature of the technology as prediction technology, noting, "Profits increase whether the rate of true positives increases at the expense of either false negatives or true negatives–a profitable sale occurs either way." Furthermore, he notes that "[t]he role of the recommen-

dation engine as advisory to the user means that its output is not the final word.... This lowers the stakes for false positives."

He also discusses voice-based user interfaces such as Alexa and Siri. These make the customer experience more efficient, but also have the power to improve workflows and replace humans in many work tasks. For example, in Agrawal et al. (2019), we describe the role of such user interfaces in replacing human transcription services in radiology. Transcription services are modular in the sense that they are easily separated from the rest of the radiologist's workflow. The costs of a mistake are relatively low in the sense that radiologists see–and can correct–mistakes quickly. The AI replaces the human task of transcription and incrementally improves productivity.

Therefore, compared to the standard task based model, the systems based intuition on modularity and stakes described in Bresnahan generates similar short-term empirical predictions on which organizations–and which units with an organization–will adopt.[1].

This paper demonstrates that a formalization of the intuition creates a distinct set of predictions, particularly around modularity and stakes, when prediction improves in ways that reduce risks and enable coordination. We provide a model of AI where tasks interact with one another and are distributed across distinct decision-makers. The purpose is to explore the barriers to the adoption of AI and to provide testable implications for how studies focusing on forecasting AI's impact should proceed. We find that Bresnahan's intuition applies in many–but not all–cases. We show that Bresnahan's emphasis on one prediction over one decision explains why modularity and low stakes lead to AI adoption. In contrast, when there can be more than one prediction over more than one decision, then less modular firms with higher stakes may benefit more from the better predictions that AI brings.

We do this using a framework that builds on Van den Steen (2017). We model a project whose payoff is the outcome of a number of decisions undertaken by different agents. If the agent knows the state, then it can take the appropriate action to maximize the stand-alone payoff. In addition to a stand-alone payoff for a good decision, there is an additional payoff if decisions are aligned. Artificial intelligence is represented by a prediction of the state given to a particular agent. Agents cannot communicate.

Using this model, we first explore the role of modularity. Consistent with Bresnahan's intuition, we find that the more organisations have distinct tasks that do not interact much, the more likely it is for AI to be adopted. Because AI involves the generation of information to inform ever more nuanced and tailored decision-making, when it is deployed in a decision

[1]Bresnahan (2020) also discusses the role of capital deepening–that firms with pre-existing useful IT capital are more likely to adopt AI. In doing so, he emphasizes that AI is a complement to IT infrastructure. While this is not a direct prediction of the task-based model, it is consistent with discussions of task-based predictions of early adoption (e.g. Webb (2020)

that is interdependent with other decisions, it can be disruptive with negative productivity implications across an entire system.

Second, this effect of modularity is compounded when AI itself is imperfect as, not only does this make relying on AI more risky, it also enhances its potential disruptive effect. This implies that, also consistent with Bresnahan's intuition, when the stakes are higher and a mistake becomes more costly, adoption is even less likely.

Third, we extend the model to examine different types of predictions. We allow for a focal state, in which coordination is likely absent additional information, and a number of other ex ante identical states. These other states are riskier, in the sense that a mistake is more costly. They are also less likely. If there is only one such state, then Bresnahan's intuition applies: that AI is more valuable in modular organizations with low stakes decisions. If, however, there are at least two riskier states, then AI can be more valuable for less modular organizations and for higher stakes. Better prediction enables useful coordination across decisions. Specifically, when improved prediction reduces the chance of picking the wrong risky state, less modular organizations benefit most, particularly when the risks of an incorrect decision are high.

Overall, this formalization of a systems-based view of the benefit of AI provides predictions of what types of organizations are likely to adopt. These predictions are different from those of the task-based model. The task-based model emphasizes modularity and a reduction in costs for prediction-based tasks. More modular organizations with high labor costs are likely to see AI adoption. The AI will replace humans in those prediction-based tasks and the organization will largely remain unchanged beyond the capital-labor substitution.

The systems-based model also predicts that AI will be adopted in modular, low-stakes situations when the AI improves prediction of the risky state. However, when a prediction is available that reduces the change of picking the wrong risky state, then the systems model generates a different prediction than the task based model. In such cases, organizations with less modularity and higher stakes benefit most because such predictions reduce the chance of a costly mistake and increase the benefit of coordination across decisions.

This idea underlies a thought experiment we highlighted in our book Prediction Machines. If Amazon's recommendation engine improves on the dimension of not recommending products that people don't want–i.e. reducing the false positives–then it makes sense to integrate the AI into the overall business model. In particular, if the AI is accurate enough about what the consumer does not want, then it is worth restructuring the organization to reduce modularity by shipping the item to the consumers door before it is ordered. Shipping would need to be integrated with the recommendation engine, eliminating the modular nature of the recommendation. A mistake is costly, as the company would need to send someone to

pick it up from the customer's home and the customer would face a hassle.

Uncertainty increasing AI versus uncertainty reducing AI Stitch Fix is uncertainty reducing – so allows easier alignment Need to put in timing to make sense of this. Which action chosen first.

2 Model Set-Up

The model here is based on Van den Steen (2017) more general in some aspects and simplified in others. Suppose that a project return, R, depends on the outcomes of K decisions, $\{D_1, ..., D_K\}$ indexed by k. Each decision results in the choice of an action, $a_k \in A_k$, a set of M_k elements. An action can be stand-alone "correct," alignment "correct," both or neither. A stand-alone correct decision, D_k, results in an increment to project return of α_k compared with an incorrect decision. If two decisions, $\{D_k, D_j\}$, are correctly aligned this results in an increment to product return of $\gamma_k j$ relative to the case where those decisions are incorrectly aligned.

The "correctness" is a decision is modelled here in a reduced form way. We suppose that the correct stand-alone decision is associated with a state, $T_k \in A_k$, knowledge of which reveals the correct decision. This state could be driven by an assessment of the external environment for a decision and/or an agent's judgment regarding the trade-offs and risks associated with particular actions. For instance, a retail manager may be considering a decision of how much to re-stock based on a prediction of future demand as well as the relative costs associated with errors in that forecast (inventory holding costs versus lost sales due to stock outs). Based on that prediction and judgement, at a given time, there is an assessment of the state and associated optimal action. If that state is correctly identified and the associated action taken, there is a boost to project return of α_k. If not, there is no such boost.[2]

We treat the correctness of the alignment decisions similarly. We suppose that whether two decisions, $\{D_k, D_j\}$, are aligned is associated with a state $T_{kj} \in \{A_k, A_j\}$. If $\{a_k, a_j\} = \theta_{kj}$, then the decision is "correct" and it contributes γ_{kj} to project return. Otherwise, there is no contribution. We assume that T_{kj} is a bijection (or one-to-one correspondence) where for every a_k there exists an action a_j that creates alignment. Thus, so long as this relationship is known, there is an alignment incentive to choose the actions that selected that state for each $\{D_k, D_j\}$. Note that the order of k and j matters here and the state θ_{kj} is different

[2]In many decisions, the "distance" from the optimal or correct decision matters. Here we abstract from those considerations but it could be imagined that α_k is a measure of the loss from deviating from the optimum and a tolerance for errors would be reflected in a relatively low α_k.

from θ_{jk}. Importantly, it may not be possible to choose both the correct stand-alone and alignment decisions in both directions. Thus, decision-makers will face a dilemma in terms of whether to choose between correct stand-alone versus alignment decisions.

That dilemma will be resolved by the relative returns to stand-alone optimality and alignment. We suppose that $R = \sum_k R_k$ where:

$$R_k = \alpha_k I_{a_k = T_k} + \sum_{j \neq k} \gamma_{kj} I_{\{a_k, a_j\} = T_{kj}}$$

where I_x are indicators that take a value of 1 if the condition, x is met and 0 otherwise. It is assumed that $\alpha_k \geq 0$ and each $\gamma_{kj} \geq 0$.

2.1 Information Structure

Van den Steen (2017), assumes that each T_k and T_{kj} are equally likely and drawn from an infinite set with $M_k \to \infty$ for all k. Thus, there is a zero probability that one agent could guess the "correct" action in the absence of knowledge of those states. Below we will consider several means by which this situation can be improved. One is that agents receive a signal of various stand-alone states for decisions they own. This may come in the form of a prediction that resolves some external uncertainty and hence, provides a clearer signal of a stand-alone state. Moving beyond Van den Steen (2017), we will consider distributions for stand-alone states that involve mass points indicating potential status-quo outcomes that agents may rely upon in equilibrium. If an agent "owns" a decision, D_k, they are the only agent who can potentially observe, in the absence of communication, the outcomes T_k and T_{kj}. Specifically, they do not know for any decision $D_{j \neq k}$, T_j nor T_{jl} including T_{jk}. Our interpretation of this knowledge set is that an agent who owns a decision is using information regarding the state as their own innate judgment regarding the trade-offs between alternative actions. In some situations, communication may make coordination possible so that, for instance, alignment actions between D_k and D_j are based on $\max\{\gamma_{jk}, \gamma_{kj}\}$ rather than one or the other.

2.2 Equilibrium Outcome with no focal actions

Suppose, for the moment, that each T_k and T_{kj} are equally likely. For instance, the probability that $a_k = T_k$ is $\frac{1}{M_k}$ and the probability that a_j is paired with a_k (i.e., $\{a_k, a_j\} = T_{kj}$ conditional on choosing a_k) is similarly $1/M_k$. In addition, suppose that states are revealed to the agent who owns D_k with probabilities p_k and p_{kj} both $> \frac{1}{M_k}$ respectively. (For states without k or where k is second in the order, there is no revelation).

Following Van den Steen (2017), we consider a case where each decision is "owned" by a distinct agent (thus, there are K agents) and they have no means of explicitly coordinating between them. Consider agent k (where we index the agent by the decision they own). Suppose agent k has complete information regarding the states they are able to observe. That agent knows that, for any given action they take, there is a $\frac{1}{M_j}$ probability that it will align with the choice of another agent j. This applies to any action including the action that maximises the stand-alone component of the agent's payoff. Thus, they expect to generate a project contribution of:

$$\alpha_k + \sum_{j \neq k} \frac{1}{M_j} \gamma_{kj}$$

By contrast, if the agent does not know the action that generates a stand-alone contribution, they can choose an action at random and expect a project contribution of:

$$\frac{1}{M_k} \alpha_k + \sum_{j \neq k} \frac{1}{M_j} \gamma_{kj}$$

Note that these outcomes do not change as any p_{kj} changes. Therefore, the expected contribution of the agent prior to receiving information (or not) about the correct stand-alone action is:

$$\frac{1}{M_k}((1 + p_k(M_k - 1))\alpha_k + \sum_{j \neq k} \frac{1}{M_j} \gamma_{kj}$$

Note that as $M_k, M_j \to \infty$ as in Van den Steen (2017), this becomes $p_k \alpha_k$. This demonstrates the difficulty of achieving the benefits from coordination if the absence of communication of some other focussing device.

2.3 The role of communication

In the model as set-up, an alignment contribution arises at random or not at all. Of course, if stand-alone actions are themselves chosen randomly, there may be no addressing this situation. However, what if the chosen stand-alone action can be communicated to agents prior to them choosing alignment actions? Suppose that such communication could occur at a cost of c per message which resulted in the receiver potentially knowing the sender's action choice. Communication is imperfect, however, so that if a sender sends a message indicating a particular action, with probability μ that message is misinterpreted as another action. In this event, alignment is not possible. Thus, if a message is sent to all decision-makers, then

the expected alignment contributions become:

$$\sum_{j \neq k} p_{kj}(1-\mu)\gamma_{kj} - (K-1)c$$

This demonstrates that the problem of alignment is a combination of the problem of judgment (knowing the state, T_{kj} so that D_j can choose the correct alignment action) and the problem of communication. When both of these are 'solved' in a given situation, which happens with probability $p_{kj}(1-\mu)$, then alignment can be achieved.

This highlights a fundamental trade-off in systems. between external responsiveness and internal alignment. If stand-alone actions were fixed and always the same, then so long as other decision-makers had judgment (and so knew T_{kj}), then alignment could be achieved. By contrast, if stand-alone actions are varying, then to achieve alignment, judgment is insufficient. Instead, communication is required. That communication both permits the possibility of alignment but also gives value to judgment.

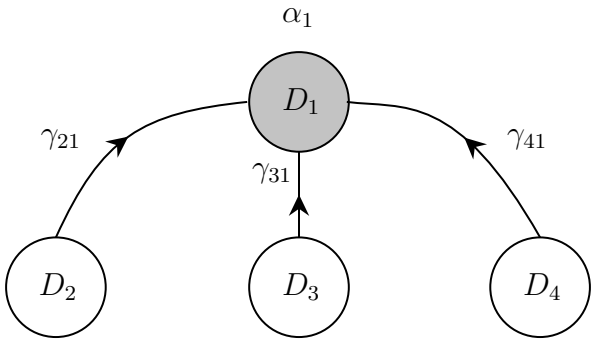
3 AI Adoption and Modularity

Our interest is in examining the impact of better prediction on how decisions are made in organisations. To this end, we make two simplifications to the model as specified above. First, the prediction we focus on is that regarding random variables external to the organisation that drive, at a first instance, the stand-alone contribution of decisions. To capture this, we turn now to examine an environment where all but one decision, D_1, has $\alpha_k = 0$. Thus, stand-alone requirements do not guide those decisions and hence, neither do predictions regarding external random variables. By contrast, for D_1, $\alpha_1 = \alpha > 0$ and thus, if better predictions enable the agent who "owns" D_1 identify the action that will be correctly matched with external uncertainty, that will impact on the likelihood that that agent's choices drive other choices in the organisation. To that end, we assume that it is only stand-alone requirements that determine 1's choices and so we set $\gamma_{1k} = 0$ for all $k \neq 1$. Figure 1 shows the resulting structure.

Second, we assume that the degree of interactions between any decision in terms of its alignment with D_1 to be symmetric, with $\gamma_{k1} = \gamma$ for all k. We also assume that the probability that an element of T_{k1} is revealed to agent k is symmetric across all alignment states with $p_{k1} = q$ for all k. Similarly, we assume that the number of actions for each decision, D_k, are also symmetric between decisions; i.e., $M_k = M$ for all k. Finally, for notational convenience, we set all other interactions γ_{kj} where $j \neq 1$ to zero.[3]

[3]Without this assumption, some additional terms are added to the equations that follow but play no role

Figure 1: **Simplified Model**



One way to represent better prediction is as an increase in p_1 which we conveniently omit the subscript. Based on the model set-up in the previous section where there are no focal actions, the marginal impact of higher p would simply be α. Without coordination, there is no means by which other decisions change as a result of better prediction of this type. In other words, the improved prediction changes the outcomes of D_1 but nothing else.[4]

Suppose, instead, that the probability distribution of T_1 is such that there exists once action, a^S which has a higher prior probability of being correct than other actions (a status quo action). Let the prior probability that a^S is the correct stand-alone action for D_1 be $\rho > \frac{1}{M}$. The probability that any particular alternative action from the remaining $M-1$ actions is optimal is $(1-\rho)\frac{1}{M-1}$.

This profoundly changes the equilibrium outcome.

Proposition 1 *(Modularity) In the simplified model, agent 1 will choose to ignore the AI signal and set $a_1 = a^S$ always, if:*

$$\alpha \leq q(K-1)\gamma$$

Otherwise, agent 1 will choose $a_1 \neq a^S$ if they receive a signal $\theta_1 = a_r \neq a^S$ and choose $a_1 = a^S$ otherwise.

Proof. First, let's conjecture that all other agents believe that $a_1 = a^S$. Then if they know T_{k1}, then will choose the appropriate correctly aligned action and earn γ. This happens with probability q. With probability $1-q$, they do not know the correct aligned action and so choose an action at random. Thus, their expected payoff is $(q + (1-q)\frac{1}{M})\gamma$. Second, if agent 1 commits to choose a^S regardless of the signal they receive, then their expected contribution is $\rho\alpha$ and the contribution of other agents is as in step 1. Third, note that if a^S is the correct stand-alone action for 1, then their expected payoff is α.

in the equilibrium outcomes.

[4]It is easy to see that this would apply to better prediction related to all external variables should other decisions have a positive stand-alone contribution, $\alpha_k > 0$.

These calculations assume that, if everyone expects agent 1 to choose a^S, the correct alignment actions to that are chosen. The only circumstance in which agent 1 would not choose a^S is where another action is known to be the optimal stand-alone action. This happens with probability $p(1-\rho)$. In this case, by choosing the alternative action, 1's contribution becomes α.

What happens to the contributions of other agents? If they were expecting 1 to choose a^S, then, if k learns T_{k1}, their realised contribution falls to 0, as they would have matched incorrectly. This happens with probability q. With probability, $1-q$, no such learning occurs. In this case, the probability that alignment is achieved is $\frac{1}{M}$. Thus, for agent $k \neq 1$, if they behave as if 1 has chosen a^S, their expected contribution is:

$$(1-p(1-\rho))(q+(1-q)\tfrac{1}{M})\gamma + p(1-\rho)(1-q)\tfrac{1}{M}\gamma$$

The alternative is for k to choose, instead, to choose an action at random which gives them an expected contribution of:

$$((1-p(1-\rho))\tfrac{1}{M} + p(1-\rho)\tfrac{1}{M})\gamma$$

Comparing these outcomes, behaving as if 1 has chosen a^S is optimal for k if:

$$1-p(1-\rho) \geq (1-p(1-\rho))\tfrac{1}{M} + p(1-\rho)\tfrac{1}{M} \implies p(1-\rho) \leq \tfrac{M-1}{M}$$

Note that the left hand side is higher the smaller ρ is. However, ρ is bounded by $\frac{1}{M}$. Substituting this in, we can see that this inequality always holds. Thus, it is always optimal for other agents to make decisions as if 1 has chosen a^S rather than another, effectively, random action.

Thus, there are two broad strategies that can be undertaken by agent 1. Let's suppose that p is sufficiently low that it is optimal for all other agents to try and align their strategies with a choice by 1 of a^S.

1. Always choose $a_1 = a^S$ which results in a total expected project return of:

$$\rho\alpha + (K-1)(q+(1-q)\tfrac{1}{M})\gamma$$

2. Choose $a_1 = a^S$ unless information is received that shows it not to be stand-alone optimal which results in a total expected project return of:

$$(p+(1-p)\rho)\alpha + (1-p(1-\rho))(K-1)(q+(1-q)\tfrac{1}{M})\gamma + p(1-\rho)(1-q)(K-1)\tfrac{1}{M}\gamma$$

Comparing these two, it is optimal for agent 1 to ignore their signal of the correct stand-alone action if $\alpha \leq q(K-1)\gamma$. ■

Proposition 1 shows that if the value of alignment is high and/or the stand-alone contribution of 1 is low, then it is optimal for agent 1 to always choose a^S. The interesting implication is that this choice is independent p – the quality of 1's information about the correct stand-alone value. In other words, improving 1's information regarding T_1, does not change 1's incentives to react to that information. This is because anything that drives 1 towards a non-focal action, causes a loss of alignment and an improvement in the expected stand-alone contribution. This can only happen when 1 learns that a^S is not optimal, with happens with probability $p(1-\rho)$. In this case, 1 is able to appropriate the stand-alone contribution by switching from a^S but that very same switch causes a loss of alignment because other agents are aligning with respect to a^S. Ironically, no such loss occurs if other agents cannot judge the precise action that would achieve alignment with a^S and instead choose their actions effectively at random. Thus, it is precisely the conscious attempt at alignment that is thwarted if agent 1 relies on the predictive signal in choosing their action.

Stability in expected choices breeds better alignment of actions from independent decision-makers. When an agent relies on prediction to determine those choices, the model here shows that the cost of this is a loss of alignment. This loss is mitigated if there are fewer interacting decisions (K) or the loss from a lack of alignment (γ) is smaller.

The implication of this for AI adoption is that, unless there is a change in how alignment is achieved in organisations, then organisations that are more modular in their design so that γ is sufficiently low, will see adoption of AI at the task level with the cost being a corresponding loss of alignment. Modularity will limit those losses. This analysis holds the means of coordinating different decisions fixed. As we explore in the next section, one of the consequences of AI adoption is a change in organisational design to accommodate it.

Before turning to that, it is useful to remark on the role of q. Recall that this is the probability that an agent $k > 1$ who owns the decision is able to learn the correct decision to align with a^S. Above we referred to this as that agent's judgment – their ability to select a decision that balances the trade-offs that might be experienced at the time to achieve alignment.

Note that as q increases, the alignment cost of adopting AI also increases. In other words, better judgment of this type is a substitute with improved prediction. That stands in contrast to other results (notably, Agrawal et al. (2018b)) who stress a complementarity here. The difference, of course, is that the prediction is with respect to one decision while the judgment is with respect to a distinct decision. However, those decisions are themselves

complements.

What is going on here is that a higher q increases the returns to choosing a^S rather than something else while better prediction, if taken advantage of, reduces the probability that a^S will be chosen and the value of that judgment is realised. If other agents had better knowledge of other actions being taken, then their judgment could be matched to those actions and improve alignment. But since such coordination is limited here, that value cannot be realised.

4 Designing for Flexible Alignment

When a decision in an organisation becomes responsive to more granular predictions about the external environment, this causes variation in the action being taken that can have implications for the choices and performance of other decisions in the organisation. Here we explore how by designing organisations to match the AI being adopted, the scope for such adoption can be enhanced. We will show that, in these cases, AI adoption requires a systemic change throughout the organisation.

4.1 Increased Modularity

If one action in an organisation starts to vary more often, how can this variation be accommodated throughout the organisation? An obvious first path would be to design the organisation to insulate the other actions from the variation arising from, say, D_1, following the adoption of AI. This could be done through a modular design that lowered γ, the degree to which decision/functions were interdependent. There are, however, costs associated with this. As Baldwin and Clark (2000) argue, apart from potential costs of loss of control of decision-makers, designing a more modular organisation is itself a potentially difficult process with significant sunk costs.

This kind of change implies that where AI has value in a system that is not otherwise modular, adopting AI will require architectural innovation that builds a new organisational structure from the ground up. This process takes time and is one of the reasons why adoption of general purpose technologies can be slow (David (1990)).

4.2 Increased Communication

Thusfar, decision-makers are assumed to act independently even if they share the organisation's goal. In reality, there may be opportunity for communication between decision-makers to facilitate alignment. For instance, should agent 1 wish to choose an action other than a^S,

a meeting or other form of communication could take place where 1's chosen action could be communicated. This would provide an opportunity for others to align their actions.

There are two constraints on communication. One is that this involves costs that are incurred each time 1 needs to communicate an alternative action. We will assume that those costs are c per message, per recipient. The other is that communication may be imperfect and there is a possibility of miscommunication whereby 1 intends to choose an action but this is interpreted by others as an alternative action. That miscommunication could be common or independent amongst recipients. We assume that every message involves a probability of miscommunication of $\mu < \frac{M-2}{M-1}$ in which case if the intended action is $\hat{a}_1$, an action $a_1 \neq \hat{a}_1, a^S$ is communicated.

Given this communication structure, suppose that agent 1 chooses to communicate an action they know that $T_1 \neq a^S$. This happens with probability $p(1-\rho)$. Given this, it is reasonable to assume that other agents believe that a^S will be chosen unless they receive a message to the contrary. If other agents receive that message, it is optimal for them to act on that message as $1-\mu > \frac{1}{M-1}$. This means that any other agent's expected alignment contribution is:

$$(1-p(1-\rho))(q+(1-q)\tfrac{1}{M})\gamma + p(1-\rho)(q(1-\mu)+(1-q)\tfrac{1}{M})\gamma - p(1-\rho)c$$

$$= (q+(1-q)\tfrac{1}{M})\gamma - p(1-\rho)(\mu\gamma + c)$$

That is, when communication takes place, three things might happen. First, with probability $q(1-\mu)$, the agent correct learns 1's choice and knows the appropriate alignment action to take. Second, with probability $q\mu$, the agent knows the appropriate alignment action to take with the action they believe 1 has chosen but that message is incorrect and so the alignment contribution is forgone. Third, with probability q, the agent does not learn the correct alignment action to take to 1's communicated action and so receives the alignment contribution with probability $\frac{1}{M}$. This final outcome does not depend on whether 1's message is garbled or not.

If 1 pursues that strategy of relying on the prediction coupled with communication of deviations from $a_1 = a^S$, the expected project return becomes:

$$(p+(1-p)\rho)\alpha + (K-1)(q+(1-q)\tfrac{1}{M})\gamma - (K-1)p(1-\rho)(q\mu\gamma + c)$$

Given the symmetry amongst those agents, whether miscommunication is common or not does not impact on this expected project return. Note that this is preferred to agent 1

choosing $a_1 = a^S$ always if:

$$\alpha \geq (K-1)(q\mu\gamma + c)$$

And this is preferred to agent 1 choosing to respond to the prediction but not communicate it if:

$$q(1-\mu)\gamma \geq c$$

And just to complete the picture, recall that agent 1 choosing to respond to the prediction but not communicate it is preferred to choosing a^S always if:

$$\alpha \geq q(K-1)\gamma$$

Note that, so long as $c \leq q(1-\mu)\gamma$, then the threshold for adopting AI is lower when communication is possible than when it is not. This happens whenever communication is preferred to not communicating.

Communication represents another way of mitigating the losses from a loss of alignment. This analysis shows that AI adoption is more likely when communication can be used in this manner. However, the simplified model here shows that communication is useful or not in its own right and the adoption of AI will not in itself drive the adoption of more communication.

5 AI Adoption and Stakes

Above we derived the conditions under which an organisation would want to pay attention to a prediction in decision-making and alter actions taken based on the signal received. It assumed that the signal, if it was received, was perfectly revealing about the true state of the world. In reality, signals are imperfect and can be biased towards false positives and false negatives regarding a state. It is the possibility of making errors that causes decision-makers to be concerned about the 'stakes' associated with their decisions. In this section, we amend the simplified model of the previous section to take this into account. The goal is to explore how interactions between decisions alter the organisation's preferences regarding the bias of predictive signals it receives.

To this end, we make the following changes to the simplified model. First, let θ_1 be a signal of T_1, the true state. Table 1 shows the signal space based on a signal of $\theta_1 = a^S$ versus two representative alternative actions a_r and a_{-r}. Action a_r represents the appropriate action in a given state $_1 = a_r$ while action a_{-r} represents an inappropriate action. This inappropriate action has a different payoff from choosing the default action a_S when the state is $T_1 = a_r$.

Row 1 of Table 1 shows that if $T_1 = a^S$, then with probability λ_S, $\theta_1 = a^S$ (i.e., a

Table 1: **Prediction Bias**

	$\theta_1 = a^S$	$\theta_1 = a_r$	$\theta_1 = a_{-r}$
$T_1 = a^S$	λ_S	$\frac{1-\lambda_S}{M-1}$	$\frac{1-\lambda_S}{M-1}$
$T_1 = a_r$	λ_{-S}	λ_r	$\frac{1-\lambda_r-\lambda_{-S}}{M-2}$

true positive) and otherwise signals another action as correct (i.e., a false negative) with probability, $\frac{1}{M-1}(1-\lambda_S)$. By contrast row 2 shows if, say, $T_1 = a_r$, it is correct with probability λ_r (i.e., a true negative relative to the status-quo) but with probability, λ_{-S}, it signals the status quo action, a^S (i.e., a false positive) and, with probability $\frac{1-\lambda_r-\lambda_{-S}}{M-2}$, it signals an alternative (non-status quo) action a_{-r} (i.e., a false negative with respect to a_r). Given this, agent 1 uses the imperfect signal to update their beliefs regarding the likelihood of various actions. In particular,

$$\Pr[T_1 = a^S | \theta_1 = a^S] = \frac{\rho\lambda_S}{\rho\lambda_S + (1-\rho)\lambda_{-S}}$$

$$\Pr[T_1 = a_r | \theta_1 = a_r] = \frac{(1-\rho)\lambda_r}{\rho\lambda_{-S} + (1-\rho)(1-\lambda_{-S})}$$

where we use the fact that $\Pr[\theta_1 = a_r] = \frac{\rho}{M-1}\lambda_{-S} + \frac{1-\rho}{M-1}(\lambda_r + \sum \frac{1-\lambda_r-\lambda_{-S}}{M-2}) = \frac{\rho\lambda_{-S}+(1-\rho)(1-\lambda_{-S})}{M-1}$. We assume, however, that these signals are informative so that $\Pr[T_1 = a^S | \theta_1 = a^S] > \rho$ and $\Pr[T_1 = a_r | \theta_1 = a_r] > \frac{1-\rho}{M-1}$ for each $a_r \neq a^S$. This places lower bounds on λ_{-S} and λ_r.

While we continue to assume that if 1 chooses a^S incorrectly, the stand-alone contribution is 0, the second change to the model is to now assume that, if 1 chooses an action other than a^S incorrectly, the stand-alone contribution is $-\beta$.[5] Thus, choosing a^S is a safer choice which carries the consequence of losing α if incorrectly selected while other actions are riskier in that an incorrect selection results not only in a loss of α but an additional loss of β. In addition, should agent 1 incorrectly choose a non-status quo action, we assume there is no ability for other agents to align with that action and generate the contribution, γ, even randomly. Thus, the risks associated with choosing an action other than the status quo are two-fold, an additional stand-alone loss and a loss of an alignment opportunity.

With this amended set of assumptions, we can show the following:

Proposition 2 *(Stakes) In the amended model, agent 1 will choose to ignore the AI signal and set $a_1 = a^S$ always, if:*

$$\beta \geq \frac{((1-\rho)\lambda_r - \rho(1-\lambda_S))\alpha + (\rho\lambda_S + (1-\rho)\lambda_{-S} - 1)(q + (1-q)\frac{1}{M})(K-1)\gamma + (1-\rho)\lambda_r(1-q)\frac{1}{M}(K-1)\gamma}{(1-\rho)(M-2)(1-\lambda_r-\lambda_{-S})}$$

[5]Note that this is similar to decision 1 having a lower α than the other decisions.

Otherwise, agent 1 will choose $a_1 = \theta_1$.

Proof. Consider a situation where agent 1 responds to the signal and chooses the action that is signaled. This is optimal by our assumption that signals are informative. As before, we start by assuming that other agents believe that agent 1 will choose a^S always and respond accordingly. In this case, the expected project return is:

$$\begin{aligned}(\rho\lambda_S + (1-\rho)\lambda_r)\alpha - (1-\rho)(M-2)(1-\lambda_r-\lambda_{-S})\beta \\ +(\rho\lambda_S + (1-\rho)\lambda_{-S})(q+(1-q)\tfrac{1}{M})(K-1)\gamma \\ +(1-\rho)\lambda_r(1-q)\tfrac{1}{M}(K-1)\gamma\end{aligned} \tag{1}$$

By contrast, recall that, if agent 1 ignores the signal and chooses a^S always, the expected project return is:

$$\rho\alpha + (q+(1-q)\tfrac{1}{M})(K-1)\gamma$$

Comparing these gives the condition in the proposition. ■

Note as signal becomes precise (that is, $\lambda_S, \lambda_r \to 1$ (and by implication $\lambda_S \to 0$) we have: $\alpha \leq q(K-1)\gamma$; the same condition as Proposition 1.

As before, note that a higher $(K-1)\gamma$ reduces the attractiveness of adopting AI. Thus, our earlier result that lower modularity restricts AI adoption continues to hold in this setting. However, we also add here the role of β. A higher β means that D_1 is a decision with higher stakes; in particular, there are greater costs associated with an incorrect decision that differs from the status quo action, a^S. This will reduce the attractiveness of adopting AI when $\lambda_{-S} > 0$. Thus, Proposition 2 reflects the intuition in Bresnahan (2020).

5.1 The type of improvement

This result suggests that as the precision of AI is improved, it will be adopted first in places with high modularity and lower stakes. However, to explore this more carefully we need to be precise about what we mean by an improvement in precision. The following proposition demonstrates how distinct improvements in AI prediction impact on the expected project return for organisations that adopt AI – i.e., ones where agent 1 relies on the AI signal.

Proposition 3 *(Precision) In the amended model, if agent 1 chooses an action based on the signal θ_1, then:*

1. *an increase in λ_r, has a marginal return of $(1-\rho)(\alpha+(M-2)\beta+(1-q)\frac{1}{M}(K-1)\gamma)$;*

2. *an increase in λ_S, has a marginal return of* $\rho(\alpha + (q + (1-q)\frac{1}{M})(K-1)\gamma)$*; and*

3. *an increase in λ_{-S}, has a marginal return of* $(1-\rho)((M-2)\beta+(q+(1-q)\frac{1}{M})(K-1)\gamma)$.

The proof simply involves taking derivatives of (1) with respect to λ_r, λ_S and λ_{-S}. The different improvements in the quality of AI prediction have different impacts on the expected return from a project. Recall that in the previous model, an increase in p, had a marginal return of $(1-\rho)(\alpha-(K-1)q\gamma)$ because this corresponded to an increase the probability that agent 1 switched their decision from the focal one, a^S, which increased the stand-alone contribution but reduced alignment with other decisions. The amended model, by contrast, expresses improvements in AI in terms of changes in the relative precision of predictions. Thus, an increase in λ_S (the prediction quality when the state is a_S), not only ensures that following the AI prediction is more likely to yield a stand-alone improvement but because the precision of that signal is higher, it is more likely to permit alignment with the focal action. That is, the higher is γ, the more valuable it is to have a clearer (or more sensitive) prediction of whether $T_1 = a^S$.

When considering changes in the other precision parameters, the level of stakes (β) plays a role. An increase in λ_r means that there is a reduced error rate in choosing an alternative action to a^S and this lower error rate is more valuable the higher is β. Interestingly, an increase in λ_{-S} which represents a higher error rate when choosing a^S when another action is optimal, also reduces the likelihood that the stakes are at put at risk so the returns to this imprecision are increasing in β. Conversely, if prediction actually becomes more accurate, reducing λ_S, higher stakes reduce the returns to such accuracy.

Finally, those other precision parameters change how the need for alignment (γ) impacts on the return to precision. In this case, an increase in λ_r is more valuable when γ is high because it increases the probability that when choosing an alternative action, alignment will be possible. An increase in λ_{-S} also has a higher return when γ is higher but this is because it reduces the chance of choosing an action other than a^S and, thus, preserves the value of aligning on that focal action.

What this implies is that as AI improves in precision on this dimension of a more reliable prediction of choosing one alternative action over another alternative action, *this will favour adoption by organisations with less modularity and more higher stakes decisions.* This analysis means we have to more clearly specify precisely how AI is improving predictions and not simply whether a prediction is becoming available or not to properly understand how organisational characteristics impact on AI adoption.

5.2 Communication Systems

To date, achieving alignment has been really on possible if the status quo action, a^S is chosen and either other agents know how to align their actions or randomly achieve alignment. Earlier we showed that communication of the prediction (or the action being taken by agent 1) could allow for further alignment. Such communication is costly, c per message and imperfect, being garbled with probability μ. However, in the context of a more precise prediction, incurring those costs could be worthwhile. Here we explore the incentives of an organisation to adopt a broader communication structure in response to more precise prediction.

If communication is possible, then it is worthwhile for agent 1 to communicate their intended action if it is something other than a^S. This arises with probability $(1-\rho)\lambda_r$. In this case, with probability $(1-\mu)q$ another agent receives that communication and has the knowledge to select the appropriate alignment action resulting in a contribution of γ. On the other hand, with probability μ, the signal is garbled. In this case, the correct alignment action is not chosen. However, if the agent does not have knowledge of the correct alignment, then they may choose the correct alignment action randomly with probability $(1-q)\frac{1}{M}$. Thus, miscommunication only is an issue if something other than a^S is being communicated.

Given all of this, with communication, (1) becomes:

$$\begin{aligned}(\rho\lambda_S + (1-\rho)\lambda_r)\alpha - (1-\rho)(M-2)(1-\lambda_r-\lambda_{-S})\beta \\ +(\rho\lambda_S + (1-\rho)\lambda_{-S})(q+(1-q)\tfrac{1}{M})(K-1)\gamma \\ +(1-\rho)\lambda_r(q(1-\mu)+(1-q)\tfrac{1}{M})(K-1)\gamma \\ -(1-\rho)\lambda_r c\end{aligned} \tag{2}$$

It can immediately be seen that this raises the return to improving λ_r to $(1-\rho)(\alpha+(M-2)\beta+(q(1-\mu)+(1-q)\frac{1}{M})(K-1)\gamma)$.

When an organisation can adopt a communication system, it only has an incentive to do so if it wants to communicate – that is, if agent 1 wants to rely on the signal θ_1 and select an action other than a^S. In this case, a decreasing modularity, that is, an increase in γ, has a marginal return of:

$$(\rho\lambda_S + (1-\rho)(\lambda_r+\lambda_{-S}))(q+(1-q)\tfrac{1}{M})(K-1) - (1-\rho)\lambda_r q\mu(K-1)$$

By contrast, without reliance on the prediction an increase in γ has a marginal return of

$(q + (1-q)\frac{1}{M})(K-1)$. Thus, increasing γ is more valuable when AI is adopted if:

$$(\rho\lambda_S + (1-\rho)(\lambda_r + \lambda_{-S}) - 1)(q + (1-q)\tfrac{1}{M}) \geq (1-\rho)\lambda_r q\mu$$

As $\rho\lambda_S + (1-\rho)(\lambda_r + \lambda_{-S}) \leq 1$ this condition does not hold. Thus, while communication mitigates the costs associated with low modularity, because that communication is imperfect, lower modularity still increases the costs associated with relying on the prediction.

6 Conclusions

This paper has provided a model of a systems based approach to understanding the impact of AI, as in Bresnahan (2020). The model provides distinct predictions about which organization will adopt AI from the task-based model that has dominated the discussion of AI in the economics literature so far. The answer depends on the nature of the predictions. If the AI predictions improve the likelihood of true positives and therefore enable a firm to take different actions from the default, then both the task based and the systems based models suggest that AI will be adopted by modular organizations for low stakes decisions. In contrast, in the systems based view, the formal model suggests that if the AI predictions reduce the likelihood of a false positive for something other than the focal action, then AI will be adopted by less modular organizations for high stakes decisions. The standard task-based view does not generate different predictions based on the type of prediction being made.

This distinction is evident in Bresnahan's primary example of marketing-focused recommendation engines. He argues that Amazon has been an early adopter of AI for its recommendation engine because it is modular and low stakes. The consequence of a true positive is helpful and a false positive matters little. However, if we allow for a different prediction tool to reduce false positives relative to increasing true positives, then such a prediction tool would be particularly valuable for non-modular organizations and high stakes decisions. That leads to the shipping-then-shipping example that we highlight in Prediction Machines.

Key to this distinction between the systems based view and the task based view is the recognition that the current generation of AI represents prediction technology, and the specifics of the type of prediction being made matters. If prediction is one-dimensional then modular and low stakes decisions will adopt AI first. If prediction is multi-dimensional, then an AI that reduces the rate of false positives is particularly likely to be transformative.

References

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Agrawal, A., Gans, J., and Goldfarb, A. (2018a). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50.

Agrawal, A. K., Gans, J. S., and Goldfarb, A. (2018b). Prediction, judgment and complexity: A theory of decision making and artificial intelligence. *The Economics of Artificial Intelligence: An Agenda*.

Aral, S., Brynjolfsson, E., and Wu, L. (2012). Three-way complementarities: Performance pay, human resource analytics, and information technology. *Management Science*, 58(2):913–931.

Baldwin, C. Y. and Clark, K. B. (2000). *Design rules: The power of modularity*, volume 1. MIT press.

Bresnahan, T. (2020). Artificial intelligence technologies and aggregate growth prospects.

Bresnahan, T., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *The Quarterly Journal of Economics*, 117(1):339–376.

Bresnahan, T. and Greenstein, S. (1996). Technical progress and co-invention in computing and in the uses of computers. *Brookings Papers on Economic Activity: Microeconomics*, pages 1–77.

Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534.

Brynjolfsson, E., Rock, D., and Mitchell, T. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers Proceedings*, 108:43–47.

Cockburn, I. M., Henderson, R., and Stern, S. (2019). The impact of artificial intelligence on innovation. *The Economics of Artificial Intelligence: An Agenda*, page 115.

David, P. A. (1990). The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American Economic Review*, 80(2):355–361.

Dranove, D., Forman, C., Goldfarb, A., and Greenstein, S. (2014). The trillion dollar conundrum: Complementarities and health information technology. *American Economic Journal: Economic Policy*, 6(4):239–70.

Felten, E. W., Raj, M., and Seamans, R. (2018). A method to link advances in artificial intelligence to occupational abilities. *AEA Papers and Proceedings*, 108:54–57.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., and Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539.

Gans, J. and Leigh, A. (2019). *Innovation+ Equality: How to Create a Future that is More Star Trek Than Terminator*. Mit Press.

Van den Steen, E. (2017). A formal theory of strategy. *Management Science*, 63(8):2616–2636.

Webb, M. (2020). The impact of artificial intelligence on the labor market. *Working paper, Stanford University*.