

Exploring the Impact of Artificial Intelligence: Prediction versus Judgment

Ajay Agrawal, Joshua S. Gans and Avi Goldfarb
February 2018

Based on recent developments in the field of artificial intelligence (AI), we examine what type of human labour will be a substitute versus a complement to emerging technologies. We argue that these recent developments reduce the costs of providing a particular set of tasks – prediction tasks. Prediction about uncertain states of the world is an input into decision-making. We show that prediction allows riskier decisions to be taken and this is its impact on observed productivity although it could also increase the variance of outcomes as well. We consider the role of human judgment in decision-making as prediction technology improves. Judgment is exercised when the objective function for a particular set of decisions cannot be described (i.e., coded). However, we demonstrate that better prediction impacts the returns to different types of judgment in opposite ways. Hence, not all human judgment will be a complement to AI. Finally, we explore what will happen when AI prediction learns to predict the judgment of humans.

1 Introduction

Artificial intelligence (AI) has advanced markedly in the past decade. With advances in machine learning – particularly deep learning and reinforcement learning – AI has conquered image recognition, language translation and game such as Go (Brynjolfsson and McAfee, 2014). This has raised the usual questions with regard to the impact of such new technologies on human productivity. Will AI substitute or complement humans in the workforce? (Autor, 2015; Markov, 2015; Acemoglu and Restrepo, 2016). In this paper, we build a simple model that takes a careful approach to precisely what new advances in AI have generated in a technological sense and applies this to a microeconomic model of task production. In so doing, we are able to provide some insight on the complements/substitutes question as well as where the dividing line human and machine performance of cognitive tasks might be. Our approach is, naturally, a first step in exploring the impact of AI but we believe sets the stage for more substantive investigations.

At the core of our approach is noting that recent developments in AI are all advances in prediction – in its statistical sense. Prediction is when you use information you do have to produce information you do not have. For instance, using past weather data to predict the weather tomorrow. Or using past classification of images with labels to predict the labels that apply to an image you are currently looking at. This is all machine learning does. It does not establish causal relationships and it must be used with care in the face of model uncertainty and limited data (Ng, 2016; Agrawal, Gans and Goldfarb, 2018a). But in an economic sense, if we were to model the impact of AI, the starting point would be a dramatic fall in the cost of providing quality predictions.

With this insight, we embed these changes in a standard model of decision making under uncertainty. As might be expected, having better predictions leads to better and more nuanced decisions – in particular, signal-contingent decisions that vary depending upon predictions received. But we note, however, that better predictions also change the returns to understanding what the payoff functions are from different (state, action) pairs. After all, if a state is so rare that you never take an action based on it, then there is no real return to understanding the payoff in that state. This process of understanding we term human judgment. At the moment, it is uniquely human as no machine can form those payoffs. This allows us to consider how better prediction interacts with the process of human judgment.

In this paper, our approach is to delve into the weeds of what is happening currently in the field of artificial intelligence (AI) to examine precisely what type of human labour will be a substitute versus a complement to emerging technologies. In Section 2, we show that prediction allows riskier decisions to be taken and this is its impact on observed productivity although it could increase the variance of outcomes as well. In Section 3, we make our stand on the issue of “what computers cannot do” and consider the role of human judgment in decision-making. Judgment is exercised when the objective function for a set of decisions cannot be described (i.e., coded). However, we demonstrate that better prediction impacts the returns to different types of judgment in opposite ways. Hence, not all human judgment will be a complement to AI. Section 4 then considers the design of prediction technology when prediction may be unreliable. Section 5 then examines span of attention issues for human judgment. Finally, in the conclusion, we explore what will happen when AI learns to predict the judgment of humans.

2 Impact of Prediction on Decisions

We now turn to model the impact of a reduction in the cost of prediction on decision-making. We assume there are two actions that might be taken: a safe action and a risky action. The safe action generates an expected payoff of S while the risky action’s payoff depends on the state of the world. If the state of the world is good then the payoff is R while if it is bad, the payoff is r . We assume that $R > S > r$. These expected payoffs represent the decision-maker’s utility from each action.

Which action should be taken depends on the prediction of how likely the good rather than the bad state will arise. To keep things simple, we will suppose that the probability of each state is $\frac{1}{2}$. Prediction is of value because it makes taking the risky action less risky. To capture this, we assume that:

$$\frac{1}{2}(R + r) < S \quad (A1)$$

Then in the absence of a prediction, the decision-maker will take the safe action. Better prediction means that the decision-maker is more likely to face a probability that is closer to 1 or 0. Thus, better prediction increases the likelihood the decision-maker receives ‘good news’ and takes the risky action.

To make this concrete, suppose that the prediction technology is such that with probability, e , the decision-maker learns that the true state. Thus, if the prediction technology is available then the decision-maker's expected payoff is:

$$\pi^m = e \left(\frac{1}{2}R + \frac{1}{2}S \right) + (1 - e)S = e\frac{1}{2}R + \left(1 - e\frac{1}{2} \right)S$$

Thus, the better is the prediction technology (e), the higher the expected payoff. The returns to creating a better prediction technology depend on $(R - S)$; the difference between the upside payoff and the safe payoff.

3 Prediction and Human Judgment

Having specified the baseline role of prediction in decision-making, we now turn to consider the application of judgment. The need for judgment arises because the decision-maker cannot describe the utility function perfectly in advance. Specifically, having received a prediction regarding the likely state of the world, the decision-maker engages in thought that allows them to assess the payoff from each action. In this mode, the payoffs specified previously are simply prior beliefs and having engaged in thought, the decision-maker can update those beliefs.

To model this, we assume that there are hidden attributes that, if known, can change the assessment of return to the risky action in the good state. Specifically, we assume that with probability $\rho/2$ ($\rho < 1$), a hidden opportunity that boosts R by Δ arises. Similarly, we assume with the same probability $\rho/2$ a hidden cost that reduces R by Δ arises. We assume that hidden opportunities and hidden costs are mutually exclusive events (i.e., they cannot both arise). Thus, the expected payoff from the risky action in the good state remains R . We focus attention on hidden attributes that are consequential. Therefore, we assume that:

$$\frac{1}{2}(R + \Delta) + \frac{1}{2}r > S \quad (\text{A2})$$

$$R - \Delta < S \quad (\text{A3})$$

Thus, if the decision-maker is uncertain regarding the state, identifying an opportunity will cause them to choose the risky action. If they are certain about the state, then identifying a cost will cause them to choose the safe action.

Judgment is the ability to recognize hidden attributes when they arise. Humans might be able to exercise judgment in identifying hidden opportunities or hidden costs and we treat their abilities in this regard as distinct. Thus, we assume that with probability, λ_g , the decision-maker gets ‘good’ news and can discover a hidden opportunity (if it arises) while with probability, λ_b , the decision-maker gets ‘bad’ news can discover a hidden cost (if it arises).¹

To begin, if there are no hidden attributes, the decision-maker will choose the safe action. If such attributes are discovered, the exercise of the resulting judgment will only change the decision if a hidden opportunity is uncovered. Therefore, the expected payoff becomes:

$$\lambda_g \left(\frac{1}{2}\rho \left(\frac{1}{2}(R + \Delta) + \frac{1}{2}r \right) + \left(1 - \frac{1}{2}\rho \right) S \right) + (1 - \lambda_g)S = \lambda_g \frac{1}{2}\rho \left(\frac{1}{2}(R + \Delta) + \frac{1}{2}r \right) + \left(1 - \lambda_g \frac{1}{2}\rho \right) S$$

The application of judgment for hidden costs does not impact on the expected payoff precisely because the discovery of such costs will not change the decision made. Notice that this opens up the possibility that a risky action will be taken in what turns out to be the bad state. Thus, judgment improves the average payoff but increases the variance.

Now consider what happens if a prediction technology (of level e) is available. In this case, the expected payoff becomes:

$$\begin{aligned} \pi^h = e & \left(\frac{1}{2} \left(\lambda_b \left(\frac{1}{2}\rho S + \frac{1}{2}\rho(R + \Delta) + (1 - \rho)R \right) + (1 - \lambda_b)R \right) + \frac{1}{2}S \right) \\ & + (1 - e) \left(\lambda_g \frac{1}{2}\rho \left(\frac{1}{2}(R + \Delta) + \frac{1}{2}r \right) + \left(1 - \lambda_g \frac{1}{2}\rho \right) S \right) \quad (1) \end{aligned}$$

Using this we can establish the following result.

Proposition 1. *Better prediction is a substitute with judgment over hidden opportunities but a complement with judgment over hidden costs.*

PROOF: The mixed partial derivative of π^h with respect to $\{e, \lambda_g\}$ is $\frac{1}{2}\rho(S - \frac{1}{2}(R + \Delta) - \frac{1}{2}r) < 0$ by A2 while the mixed partial derivative of π^h with respect to $\{e, \lambda_b\}$ is $\frac{1}{2}\rho(S - (R - \Delta)) > 0$ by A3.

The intuition is simple. Without prediction, only ‘good news’ will change the decision from the safe default. As prediction becomes better, then the decision-maker is more likely to choose the

¹ In Agrawal, Gans and Goldfarb (2018b), we provide a model with a more explicit view of how judgment leads to the formation of payoff functions. The cost of judgment is explicitly the time cost of engaging in thought about what the value of (state, action) pairs are. In Agrawal, Gans and Goldfarb (2018c), we extend this approach to consider judgment as arising from experience with additional dynamic consequences.

riskier action. However, in this situation, it is only ‘bad news’ that will cause the decision-maker to revert to the safe action. In other words, judgment is useful when it changes a decision from that which would be determined by information about the uncertain action. When there is little information (i.e. without prediction) only judgment on hidden opportunities can change the decision. When there is good information (i.e., prediction is more precise) only judgment on hidden costs can change the decision. The outcome here is related to the ‘bad news principle’ that arises in decisions regarding the timing of irreversible investments (Bernanke, 1983). In that situation, the option value of delaying an investment is only positive if there is information that can be gathered that causes the investment to be abandoned.

It is useful to note that when judgment is generic (i.e., that the exercise of judgment is equally applicable for hidden opportunities and costs, $\lambda_g = \lambda_b$), then prediction and judgment are complements only if $S - r > 2(R - S)$; that is, if the downside risk is much larger than the upside risk.

4 Unreliable Prediction

Thus far, when the machine returns a prediction, it delivers a perfect signal regarding whether the good or bad state is true or not. But what if the prediction itself is imperfect? Suppose that the confidence that the prediction is true is $a < 1$ meaning that the probability of a false positive is $1 - a$. We assume that this confidence is independent of human judgment. In particular, even if there is a false positive, the probability of a hidden opportunity or hidden cost is unchanged. Finally, we assume that the prediction is sufficiently reliable so that $aR + (1 - a)r > S$. We focus here on the unreliability of a prediction that the state is good. As it turns out, that since the signal that a state is bad will not change a decision to choose S , the reliability of that prediction does not matter.

Given this change we can now write π^m and π^h as:

$$\begin{aligned}\pi^m &= e \left(\frac{1}{2}(aR + (1 - a)r) + \frac{1}{2}S \right) + (1 - e)S \\ \pi^h &= e \left(\frac{1}{2} \left(a \left(\lambda_b \left(\frac{1}{2}\rho S + \frac{1}{2}\rho\Delta + \left(1 - \frac{1}{2}\rho \right) R \right) + (1 - \lambda_b)R \right) + (1 - a)r \right) + \frac{1}{2}S \right) \\ &\quad + (1 - e) \left(\lambda_g \frac{1}{2}\rho \left(\frac{1}{2}(R + \Delta) + \frac{1}{2}r \right) + \left(1 - \lambda_g \frac{1}{2}\rho \right) S \right)\end{aligned}$$

Obviously, given a choice, one would want a more reliable prediction. Here, because reliability does not alter the decision in the absence of judgment, the level of reliability does not impact on the association between prediction and judgment. That is, Proposition 1 still qualitatively holds.

Instead, the interesting question is a *design* one: suppose it was the case that if you want a prediction to be reported more often (a higher e), then that only comes about with a sacrifice in reliability (a). That is, you can design the machine prediction technology to be more optimistic (that is, reporting a prediction that the state is positive more often) but at the expense of that prediction being true less often. By contrast, a cautious prediction would be one that it was reported more sparingly but that was more likely to be true when reported. An alternative interpretation of this trade-off is to consider e as not simply a prediction but the ability of a human to parse the prediction (that is, to understand it). In this interpretation, the more a prediction can be explained, the less reliable it becomes. Regardless of interpretation, what interests us here are situations where there is a technical constraint that relates the reliability of prediction to its availability.

To consider this, assume that the technical relationship between e and a is described by $e(a)$, a decreasing, quasi-concave function. What we are interested in is how the effectiveness of human judgment (in particular, λ_b) changes the type of prediction technology chosen.

Proposition 2. *Suppose that $R - r \leq \Delta$, then as λ_b increases, the optimal value of e increases while the optimal value of a decreases.*

PROOF: The equilibrium point is where the slope of $e(a)$ equals the marginal rate of substitution between e and a ; that is,

$$\frac{\partial \pi^h / \partial e}{\partial \pi^h / \partial a} = \frac{\left(\frac{1}{2}(a(\lambda_b \frac{1}{2} \rho(S + \Delta) + (1 - \lambda_b \frac{1}{2} \rho)R) + (1 - a)r) + \frac{1}{2}S\right) - (\lambda_g \frac{1}{4} \rho(R + \Delta + r) + (1 - \lambda_g \frac{1}{2} \rho)S)}{e \left(\frac{1}{2}(\lambda_b \frac{1}{2} \rho(S + \Delta) + (1 - \lambda_b \frac{1}{2} \rho)R - r)\right)}$$

The sign of the derivative of this with respect to λ_b is the same as the sign of:

$$2 \left(2S(S + \Delta) - R(S + a(S + \Delta)) + r(R - (2 - a)(S + \Delta)) \right) - \lambda_g(r + R - 2S + \Delta)(R - 2(S + \Delta))\rho$$

Taking the derivative with respect to a :

$$2(S + \Delta)(r - R) < 0$$

Thus, it is decreasing in a . The highest a can be is 1 in which case the expression becomes:

$$2((2S - r)(S - R + \Delta) - R\Delta) - \lambda_g(r + R - 2S + \Delta)(R - 2(S + \Delta))\rho$$

The second term is positive by our earlier assumptions. The first term is positive if

$$(2S - r)(S - R + \Delta) > R\Delta \Rightarrow (2S - r)(S - R) > (R + r - 2S)\Delta$$

which cannot hold. However, setting $\lambda_g \rho = 1$, we can show that the overall expression is positive if:

$$2\Delta(S + \Delta) > R(R + \Delta - r)$$

This holds so long as $R - r \leq \Delta$. The lowest a can be is $\frac{S-r}{R-r}$ which, when substituted into the expression, becomes:

$$-2(r - S)(-R + S + \Delta) - \lambda_g(r + R - 2S + \Delta)(R - 2(S + \Delta))\rho$$

which is positive.

Intuitively, when R is relatively low, the consequences of unreliability are relatively high but the application of judgment serves to protect against the consequence of that unreliability (namely, choosing the safe action). Thus, in designing the prediction technology, better judgment favors choosing to have prediction under more circumstances but with lower reliability.

When R is above the threshold in Proposition 2, a clear monotone comparative static does not arise. While an increase in λ_b always serves to mitigate the consequences of unreliability, it is also a complement with prediction itself. When the consequences of unreliability are relatively low, it may be that the strength of complementarity between λ_b and prediction outweighs it, causing the decision-maker to adjust towards more prediction even if it is less reliable.

5 Inattention and Real Machine Authority

Thus far, we have considered the roles of prediction and judgment in the context of a single decision. However, individuals usually have a number of decisions that are under their span of control. Moreover, those decisions may differ in terms of their underlying drivers of the costs of prediction and the value of judgment. In this section, we ask when an individual who has formal authority for a decision may, in effect, delegate that decision to a machine by choosing not to pay attention and exercise judgment.

This question is related to the study of Aghion and Tirole (1997) on formal versus real authority. In their context, a subordinate who collected information – similar to our notion of prediction here – would report that information to a superior who could also collect their own information. Because the superior could not pay attention to everything, the subordinate might be able to exercise real authority over some decisions. Moreover, because the subordinate may gain from that authority privately, they would have a greater incentive to collect information.

Here we need not concern ourselves with a conflicted or unmotivated subordinate. The machine will engage in prediction regardless and there is little reason to suppose that engaging in prediction over a wider domain will reduce the quality of their predictions. However, the individual who has formal authority over the decision may face such issues. That individual may wish to exercise judgment but in some circumstances, paying attention to a larger number of factors may limit their effectiveness in finding judgment opportunities. Thus, it becomes important where that individuals focusses their attention and how this focus changes with the underlying environment.

To consider this, we amend the baseline model (where $a = 1$) as follows. There is now a continuum of environments the decision-maker might be faced with each with a different Δ . Each Δ comes from the $[\underline{\Delta}, \bar{\Delta}]$ domain and has a frequency $f(\Delta)$, where $\int_{\underline{\Delta}}^{\bar{\Delta}} f(\Delta) d\Delta = 1$ and $F(\cdot)$ is the cdf of $f(\cdot)$. The lower bound is such that our earlier assumptions (A2) and (A3) are maintained. Otherwise, all parameters are held constant.²

We imagine that the decision-maker can choose the number of environments they monitor. If they do not monitor the environment, they cannot exercise judgment. In that situation, their expected payoff is π^m as stated above. In this situation, the prediction of the machine determines fully the decision taken. Hence, we say that in this case the machine has real authority. If they monitor an environment, they can exercise judgment of the form we described in the baseline model for that environment. In that case, their expected payoff is as listed in (1).

Note that the expected payoff in (1), π^h , is greater than that in (2), π^m , and it is increasing in Δ . If there were no costs to doing so, therefore, the decision-maker would monitor all environments and exercise judgment in them as necessary. To consider such costs, let D be the set of environments that are monitored by the decision-maker. We will assume that $\lambda_b(|D|)$ is decreasing in $|D|$, the cardinality of D . The idea is that the fewer environments a decision-maker monitors, the better they become at exercising judgment in any one.

Given this we can demonstrate the following:

² The structure of this model is similar to that of Athey et.al. (1994) and the results that follow have their analogs in their findings. Where there are differences is that we only allow an expansion of effect decision-authority to negatively impact on the quality of one-side of the equation (human judgment) whereas machine prediction is assumed to be scalable without diminishing returns. In addition, we side-step an issue with their model regarding the incentives of different agents by considering a setting in which there are no apparent conflicts of interest.

Proposition 3. *There exists a cut-off environment, Δ^* , so that the machine has real authority over environments $[\underline{\Delta}, \Delta^*]$ and the human exercises judgment over environments $[\Delta^*, \bar{\Delta}]$. At the optimal span of control Δ^* , $\pi^h(\Delta^*, \lambda_b(\bar{\Delta} - \Delta^*)) \gg \pi^m$.*

The proposition says that the human exercises judgment in the states with the highest Δ – a parameter that we noted captured the value of judgment. But more importantly, the human cedes real authority to the machine even in states where it has an absolute advantage contingent upon its equilibrium level of judgment ability. The reason this occurs is that the human finds it optimal to take a ‘hands off’ approach in environments where judgment has lower value so that they can improve their monitoring in higher value states. In effect, they specialize their attention. Observationally, there exist environments where the human could have exercised judgment and taken a safer course of action but does not.

We now turn to examine how changes in the external environment impact on the degree of real machine authority. To begin, suppose that prediction technology improves (i.e, an exogenous increase in e). We can prove the following:

Proposition 4. *Δ^* is increasing in e if $\lambda_b(S + d - R) \leq \lambda_g(R + d + r - 2S)$.*

PROOF: The problem being solved is:

$$\Delta^* = \operatorname{argmax}_d \int_{\underline{\Delta}}^d \pi^m(e) dF(\Delta) + \int_d^{\bar{\Delta}} \pi^h(e, d, \Delta) dF(\Delta)$$

The proposition will hold if the objective function is supermodular in (e, d) . The derivative of the objective function with respect to d is:

$$\pi^m(e)f(d) + \int_d^{\bar{\Delta}} \frac{\partial \pi^h(e, d, \Delta)}{\partial d} dF(\Delta) - \pi^h(e, d, d)f(d)$$

Taking the derivative of this with respect to e gives:

$$\frac{\partial \pi^m(e)}{\partial e} f(d) + \int_d^{\bar{\Delta}} \frac{\partial \pi^h(e, d, \Delta)}{\partial d \partial e} dF(\Delta) - \frac{\partial \pi^h(e, d, d)}{\partial e} f(d)$$

$$\frac{\partial \pi^h(e, d, \Delta)}{\partial d \partial e} = \frac{\partial \lambda_b \rho}{\partial d} \frac{1}{4} (S + \Delta - R) > 0$$

$$\frac{\partial \pi^m(e)}{\partial e} = \frac{1}{2} (R - S) > 0$$

$$\frac{\partial \pi^h(e, d, d)}{\partial e} = \lambda_b \frac{1}{4} \rho (S + d - R) - \lambda_g \frac{1}{4} \rho (R + d + r - 2S) + \frac{1}{2} (R - S) > 0$$

Note that:

$$\frac{\partial \pi^h(e, d, d)}{\partial e} \leq \frac{\partial \pi^m(e)}{\partial e} \implies \lambda_b(S + d - R) \leq \lambda_g(R + d + r - 2S)$$

which is the condition in the proposition.

The first condition in the proposition says that the difference $\pi^h(e, d, d) - \pi^m(e)$ is non-increasing in e ; that is, as the prediction technology rises, the marginal benefit to the area where judgment is not applied increases by more than that where it is applied. However, this proposition provides a sufficient condition only. For a necessary and sufficient condition, if we assume that f is uniform so that $f(\Delta) = \frac{1}{\Delta - \bar{\Delta}}$. Then Δ^* is increasing in e if and only if:

$$\frac{\partial \lambda_b}{\partial d} \frac{1}{2} (\bar{\Delta} + d - 2R + 2S)(\bar{\Delta} - d) + \lambda_g(R + d + r - 2S) - \lambda_b(S + d - R) \geq 0$$

There are two broad effects. The first is the strength of complementarity between generic judgment and prediction. The stronger this is, the less likely it is for this inequality to hold. The second is the degree of reduction as the span of control increases. The larger this is, the more likely it is for the inequality to hold. Intuitively, therefore, prediction improvements will be associated with an increase in machine real authority if those improvements mean that it pays to improve judgment on inframarginal units rather than expand judgment at the margin.

As a final comparative static, let us consider a change in the underlying complexity of the environment. Suppose that $f(\Delta; \theta): [\Delta, \bar{\Delta}] \times \mathbb{R} \rightarrow [0, 1]$ is our density function now parameterized by θ . Suppose that the likelihood ratio, $f(\Delta; \theta)/f(\Delta; \theta')$ is monotone increasing in Δ for $\theta > \theta'$. This means that an increase in θ shifts the mass of the density so that highly ordered states occur relatively more frequently. This could be interpreted as an increase in complexity where the effect of human judgment is more likely to be consequential in higher than lower ordered environments.

Given this we can demonstrate the following:

Proposition 5. Δ^* is increasing in θ .

The proof follows Athey et.al. (1994) and is omitted. Intuitively, as states where the human is likely to exercise their formal authority occur relatively more frequently, the benefits to them having superior judgment in those states increase. Therefore, they reduce their span of control, in order to concentrate better judgment across fewer environments.

This result is important because it suggests that, in contrast to Acemoglu and Restrepo (2016), an increase in complexity may increase the range of environments that machines exercise real authority and reduce those for humans. That said, it cannot be said whether the overall intensity of human judgment application increases or decreases as a result of this change. That is, humans exercise judgment more often per environment but the set of environments where they do so has been reduced.

6 Conclusions

In this paper, we explore the consequences of recent improvements in machine learning technology that have advanced the broader field of artificial intelligence. In particular, we argue that these advances in the ability of machines to conduct mental tasks are driven by improvements in machine prediction. Therefore, we examine sources of comparative advantage in the presence of improved machine prediction. Specifically, we identify judgment as an activity, distinct from prediction, that can improve decision-making outcomes. Whereas prediction is information regarding the expected state of the world that can be easily described, judgment relies on factors that are indescribable. These are things often classed as intuition, transference, and the drawing of analogies for unfamiliar situations. To be sure, judgment is not a passive activity. Instead, it requires deliberative cognitive application.

With this distinction formalized, we demonstrate several outcomes. The first is that when it comes to whether more prediction enhances the value of judgment or not, the type of judgment matters. This is because prediction tends to favor choosing actions that are riskier when undertaken without information. So, while, in the absence of prediction, judgment might push decisions towards riskier actions if it identifies hidden opportunities (that is, factors that make the riskier action have greater up-side potential), when there is prediction, the reverse is true. In such situations, the role of judgment is to move away from the riskier action to safer ones. And this will happen if that judgment is directed at identifying hidden costs associated with the riskier action. Thus, we argue that as the cost of prediction falls and prediction is applied to more decisions, then the type of judgment that is valuable will move from judgment regarding good news to judgment regarding bad news.

We then turn to consider trade-offs in the design of prediction technology. In this regard, we argue that there is a potential trade-off between more predictions being generated and made available to decision-makers and the reliability of those predictions. The notion is that a machine might signal a favorable state to a decision-maker more often but the cost of this is that any given prediction is more likely to be incorrect. We demonstrate that it can often be the case that better judgment (regarding bad outcomes) will tend to push the design of prediction technology towards choosing more situations where prediction is used, at the expense of those predictions being less reliable. This reinforces the notion that as the costs of machine prediction fall, we will likely see more variance in actual outcomes than previously even if such variance is associated with higher average returns.

Finally, we examine span of control issues for humans. In our model, a human has formal authority over any decision. However, if a decision is determined exclusively by a machine's prediction, then that authority may be abrogated; that is, the machine may have real authority. Why might this occur? We demonstrate that the allocation of real decision authority towards machines arises because, in a diverse environment, humans may only monitor a limited number of contexts with which to apply judgment as doing more reduces the quality of judgments on infra-marginal environments. We show that this implies that humans will cede authority even in environments where they might apply judgment (at its current quality) efficiently. Moreover, we demonstrate that as the frequency of environments where judgment is more valuable increases (something that may be interpreted as an increase in complexity), then the human will cede an even greater number of environments to machine real authority.

Our analysis, however, remains just the beginning and there are numerous directions in which our exploration of prediction and judgment can proceed. For instance, thusfar, we have defined judgment in the negative: an intervention that cannot be described. This is necessary because if it can be described then it is conceivable that it could form part of the prediction that machines can undertake. In reality, this is useful only in a static context. One of the hallmarks of machine learning is that prediction can improve over time as the machines are able to observe the outcomes associated with an observed state and a chosen action. Thus, for instance, when a human intervenes because they assess there are hidden costs, this intervention becomes data for the machine. Given enough such interventions, the machine can potentially form an inference regarding when those costs arise. Hence, the machine would be less likely to recommend the risky

action if similar circumstances arise in the future. From our perspective, at the margin, this new data would then allow prediction to replace judgment.

Machines could even be designed so that the process of having prediction replace judgment is deliberate. To get machines to learn judgment, the machines need examples of judgment; so they can observe it in more environments and learn to mimic it. This may mean you want the machine to return predictions more often to encourage humans to consider alternatives. Of course, you might always want to withhold prediction and observe judgment so as to observe and learn from ‘good news’ judgment as well. The dynamics here are likely to be subtle. Furthermore, this suggests a potential limit to the ability of machines to learn judgment: The need for a willing human trainer. Thus, within organizations, human workers have incentives to sabotage the training process if the purpose of training is for the machines to replace the human workers.

It is useful to consider other limits on the potential for judgment to become prediction over time. In addition to the within organization issues around incentives and sabotage, two other constraints on judgment becoming prediction are rare events and privacy. Both arise because of the inability of the machine to collect enough data to make a prediction. For rare events, the machine may never get enough observations to learn even if it observes the human judgment each time. Thus, it is difficult to predict presidential elections because, by definition, they happen only every four years. Privacy concerns limit the ability of the machines to collect data. Without data, predictions are not as accurate and judgment will be needed to fill in the missing information. Thus privacy, as manifested in refusal of some humans to provide certain information to machines, generates an important role for judgment (in the form modeled above).

In contrast to this discussion about the limits of machine prediction to replace human judgment, an alternative viewpoint is that human judgment is itself deeply flawed and so having less of it will improve decision-making in many circumstances. As behavioral economists and psychologists have shown us, human judgment has biases. In the model, judgment is accurate. This raises the interesting question of whether improved machine prediction can counter such biases or might possibly end up exacerbating them.

7 References

- Acemoglu, Daron and Pascual Restrepo. 2016. The Race Between Machine and Man: Implications of Technology for Growth, Factor Shares, and Employment. Working paper, MIT.
- Aghion, Philippe, and Jean Tirole. 1997. Formal and real authority in organizations. *Journal of political economy*, 1-29.
- Agrawal, A., J.S. Gans and A. Goldfarb (2018a), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review Press: Boston.
- Agrawal, A., J.S. Gans and A. Goldfarb (2018b), “Prediction, Judgment and Complexity,” in Agrawal et.al. (eds.), *The Economics of Artificial Intelligence*, NBER/University of Chicago Press.
- Agrawal, A., J.S. Gans and A. Goldfarb (2018c), “Human Judgment and AI Pricing,” *American Economic Association Papers and Proceedings*, forthcoming.
- Athey, S., J.S. Gans, S. Schaefer and S. Stern. 1994. The Allocation of Decisions in Organizations. *unpublished manuscript*, Stanford.
- Autor, David. 2015. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* 29(3), 3-30.
- Bernanke, Ben S. 1983. Irreversibility, Uncertainty, and Cyclical Investment. *The Quarterly Journal of Economics* 98(1): 85-106.
- Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age*. W.W. Norton, New York.
- Domingos, Pedro. 2015. *The Master Algorithm*. Basic Books, New York.
- Markov, John. 2015. *Machines of Loving Grace*. HarperCollins Publishers, New York.
- Ng, Andrew. 2016. What Artificial Intelligence Can and Can’t Do Right Now. *Harvard Business Review Online*. <https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now>. Accessed December 8 2016.