NBER WORKING PAPER SERIES

THE ALLOCATION OF DECISION AUTHORITY TO HUMAN AND ARTIFICIAL INTELLIGENCE

Susan C. Athey
Kevin A. Bryan
Joshua S. Gans

Working Paper 26673
http://www.nber.org/papers/w26673

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2020

Thanks to Jorge Guzman for an excellent discussion. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w26673.ack

The Allocation of Decision Authority to Human and Artificial Intelligence
Susan C. Athey, Kevin A. Bryan, and Joshua S. Gans
NBER Working Paper No. 26673
January 2020
JEL No. C7,M54,O32,O33

ABSTRACT

The allocation of decision authority by a principal to either a human agent or an artificial intelligence (AI) is examined. The principal trades off an AI's more aligned choice with the need to motivate the human agent to expend effort in learning choice payoffs. When agent effort is desired, it is shown that the principal is more likely to give that agent decision authority, reduce investment in AI reliability and adopt an AI that may be biased. Organizational design considerations are likely to impact on how AI's are trained.

Susan C. Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Kevin A. Bryan
University of Toronto
Canada
kevincure@gmail.com

Joshua S. Gans
Rotman School of Management
University of Toronto
105 St. George Street
Toronto ON M5S 3E6
CANADA
and NBER
joshua.gans@gmail.com

Artificial intelligence (or AI) adoption is often equated with automation, whereby humans are replaced by machines in tasks and decisions. In practice, however, AI is commonly used to augment human activity. Consider partially self-driving cars with human override, suggested scripts for customer service, and scoring for risk or priority in hiring, audits, judicial sentencing and fraud detection. Decisions often involve considerations that are difficult to digitize or where prior knowledge is important for anticipating outcomes in novel or unusual circumstances. These are areas where the automated predictions of fully automated AI can be insufficient, even when AI reduces the cost of prediction along some margins (Agrawal et al. [2018]). This motivates an analysis of precisely how humans and AI would work together and, as we will focus on here, under what circumstances the human or the AI should have formal decision-making authority.

On the face of it, AIs have several characteristics that make them desirable decision-makers. Classic incentive theory weighs the benefits of high-powered incentives against the risk they impose on risk averse agents; AI are not, or at least can be programmed not to be, risk averse. They do not have significant 'effort' costs beyond the fixed costs associated with developing them. And, as some have argued, they are easier to control and act in the interests of a principal. On the other hand, there remains the challenge of providing an AI with the 'correct' objective function, which requires defining and digitizing principal objectives.[1] On the other hand, human decision-making may be more accurate than AI choices in some circumstances, especially when there are deficiencies in the data available to train and operate an AI.

Consider two stylized cases: a self-driving long haul truck, and an AI that helps filter job applicants. The principal and the human agent - a driver or hiring manager - may disagree about optimal actions. A human driver may want to drive too fast on the highway, or route via their friend's town for lunch. A hiring manager may want to discriminate on the basis of race or gender. If the AI can always make principal-optimal decisions for any realization of new data, it is always optimal to ignore the human agent. However, when the AI makes mistakes - either because its training is biased or because some relevant information is not digitalizable - it may be better to combine the AI and the agent's knowledge in some way. For example, a human driver may be better able to handle unusual construction-related street patterns, and a hiring manager may be able to observe characteristics of an applicant beyond their CV.

These differences motivate us to examine organizational design choices that involve a human agent and an AI working together on a task or decision. While there are many relevant organizational design elements, including incentives, job design, and communication

[1] See Hadfield-Menell and Hadfield [2019]

patterns, we focus here on the pure allocation of decision-authority. Using the canonical model of authority from Aghion and Tirole [1997] altered to focus on human-AI interactions, we consider a principal who faces a choices as to whether to give a human agent or an AI authority in making a decision. How does the introduction of the AI affect human effort? When AIs predict well, might humans decrease effort too much ("fall asleep at the wheel")? When should the AI or the human have the right to make the final decision? Are "better" AIs in a statistical prediction sense necessarily more profitable for an organization?

While others have examined the implementation of AI in organizations,[2] this is the first paper that focuses explicitly on the interaction of control problems for humans versus AI. Importantly, we identify when organizations may prefer to hold back their investments in AI performance.

1 Model Set-up

The initial model setup follows Aghion and Tirole [1997] where there is a principal (P) who allocations decision authority and a (human) agent H who expends effort in learning information about the (expected) value of a set of projects. The projects have payoffs to P and H respectively of $(\alpha B, b), (B, \beta b)$ and $(-K_P, -K_H)$ but which project has which payoff is, initially, unknown. We assume that both α and β, the 'congruence parameters,' lie on $(0, 1]$ making the first and second projects agent and principal preferred respectively. In addition, there exists a neutral project with payoff normalized to $(0, 0)$. As in Aghion and Tirole [1997], it is assumed that $(-K_P, -K_H)$ is sufficiently negative that both P and H would prefer the neutral (or no implementation) choice over a blind choice over all projects.

Initially, the agent does not know any project's value, but following Aghion and Tirole [1997], can select effort e at cost $g(e)$, and thus learn the agent's payoff for *all* projects with probability e. We assume disutility of effort is increasing, strictly convex, $g(0) = 0, g'(0) = 0$, and $g'(1) = \infty$. Note that, absent other information or decision-makers, when an agent learns project payoffs and selects their preferred project, the principal prefers that choice to the neutral project.

Similarly, if the principal has an AI of capability E available, we assume that AI can, without any additional cost, learn the value of *all* projects with probability E. In this case, the AI will be able to select (or communicate costlessly to P) which project is the principal's preferred project.[3] E is assumed to be common knowledge.

[2] For instance, Agrawal et al. [2019], Dogan and Yildirim [2017], Dogan et al. [2018] and Aghion et al. [2019]

[3] In Aghion and Tirole [1997], P was assumed to have the ability to learn project values with probability E provided they incurred an effort cost. Here, we have endowed P with an AI that can learn on their behalf

We consider the following timing. First, decision rights are allocated: either the AI or the agent is delegated formal final decision authority.[4] Second, the agent chooses how much effort to exert. Third, the non-delegated player then reports any subset of project payoffs to the delegated player, where this report is verifiable. Finally, the delegated player chooses a project.

We also assume that the participation constraint for the agent is never violated; that is, we consider only how decision rights affect the agent's intensive margin of effort searching for projects. Letting the agent outside option be 0 suffices.

2 Allocating Decision Authority

The principal's choice regarding whether to give the agent or the AI decision authority depends on their payoff in anticipation of the agent's choice of effort in learning about project payoffs. If the agent holds decision rights, payoffs are as follows:

$$u_P = e\alpha B + (1-e)EB$$

$$u_H = eb + (1-e)E\beta b - g(e)$$

That is, the agent learns the principal's preferred project with probability e and implements it. Otherwise, the agent accepts the AI's preferred action if the AI makes a recommendation, and implements the neutral action otherwise. If the AI holds decision rights, these payoffs become:

$$u_P = EB + (1-E)e\alpha B$$

$$u_H = E\beta b + (1-E)eb - g(e)$$

In this case, if the AI learns the payoffs, the principal will implement the project they prefer, otherwise, if only the agent learns the payoffs, the principal will accept the agent's recommended project.

Let $\hat{e}_H$ and $\hat{e}_{AI}$ be the agent's effort choices under its own (human) authority and the AI's authority respectively. These are determined by the following first order conditions:

$$(1-E\beta)b = g'(\hat{e}_H) \tag{1}$$

$$(1-E)b = g'(\hat{e}_{AI}) \tag{2}$$

and H knows the AI's capabilities.

[4]Note that decision rights can be conditional on the parameters commonly observed.

A comparison of these conditions shows that the human's marginal benefit of learning is higher when they hold decision rights (as $\beta \leq 1$) so that $\hat{e}_H \geq \hat{e}_{AI}$. This formalizes a cost of delegating to an AI: when the AI has decision rights, the agent is tempted to "fall asleep at the wheel" since the AI frequently makes the choices. Even when the agent has decision rights, if the AI is an attractive "backstop" (that is, the AI's recommended project is more aligned with the agent as β increases), then the agent also has reduced incentives for effort. Finally, agent effort is decreasing in the 'quality' (E) of the AI; that is, they are strategic substitutes.

Given this, P will choose to give the AI (rather than H) decision authority if:

$$\frac{1-E}{1-E\frac{1}{\alpha}} \geq \frac{\hat{e_H}}{\hat{e_A}}$$

As the RHS exceeds 1, AI authority is only optimal if α is sufficiently low. If $E \geq \alpha$, AI authority is always optimal; the human is so misaligned that even arbitrary human effort provision due to delegation is less profitable than the imperfect AI making final decisions.

Using a specific functional form for $g(e) = \frac{1}{1+\gamma}e^{1+\gamma}$ with $\gamma > 0$, the condition for favoring the AI becomes:

$$\frac{1-E}{1-E\frac{1}{\alpha}} \geq (\frac{1-E}{1-E\beta})^{\gamma}$$

which as $\gamma \to 1$, becomes $1 \geq \alpha\beta$ (assuming that $E < \alpha$). Thus, it is optimal to allocate decision authority to the AI, so long as γ, a driver of the effort elasticity of cost, is not too high. Thus, in determining whether to give the AI decision authority, P will weigh the potentially greater reliability of the AI in selecting projects against the difficulty of motivating H to expend more effort to identify projects with non-negative returns for P.

3 The Demand for AI Performance

Given that, in the model thus far, there is no cost to P in developing an AI with higher performance, E, it is natural to presume that only technical constraints would limit the level of E employed. However, as the above analysis shows, when the probability that the AI learns project payoffs increases, the effort expended by the human agent falls. This reduces the payoff to P as it reduces their payoff in the scenarios where the AI does not learn project payoffs. Does this possibility imply that P might choose to deploy an AI with performance below what is technically feasible?

To answer this question, we begin by identifying the conditions under which P utility will be non-decreasing in E for all E.

Proposition 1 *The principal will always prefer an AI with higher E if (i) α is sufficiently low or (ii) $|\frac{g'(\hat{e}_{AI})}{g''(\hat{e}_{AI})(1-\hat{e}_{AI})}| \leq 1$.*

The proof is as follows. The derivative of P utility in E when the AI has decision authority is:

$$\frac{du_P}{dE} = (1-\hat{e}_{AI}\alpha)B + \frac{d\hat{e}_{AI}}{dE}\alpha(1-E)B$$

and when the agent has decision rights is

$$\frac{du_P}{dE} = (1-\hat{e}_H)B + \frac{d\hat{e}_H}{dE}(\alpha - E)B$$

When are these derivatives non-negative? Note first that for α close to 0 each of these derivatives are positive as $\hat{e}$ is independent of α. Rearranging terms and assuming that $\alpha > E$, we need $|\frac{d\hat{e}_{AI}}{dE}| \leq \frac{\frac{1}{\alpha}-e}{1-E}$ and $|\frac{d\hat{e}_H}{dE}| \leq \frac{1-e}{\alpha-E}$. As $\alpha \to 1$, both inequalities collapse to $|\frac{d\hat{e}}{dE}| \leq \frac{1-e}{1-E}$. That is, as $E \to 0$, we need $\frac{d\hat{e}}{dE} \to 0$: H effort needs to decrease arbitrarily slowly in E when E is low. Note that $|\frac{d\hat{e}_AI}{dE}| = |\frac{b}{g''}| > |\frac{d\hat{e}_H}{dE}| = \beta|\frac{b}{g''}|$ and under AI authority, $b = \frac{g'(\hat{e}_{AI})}{1-E}$ meaning that with substitutions the condition becomes (ii).

This shows that so long as (i) P and H are not sufficiently aligned in their project preferences or (ii) the responsiveness of H effort to improvements in E is not too great, then the impact of better AI on the incentives of the agent will not outweigh the benefits P receives from employing that AI. Significantly, the analysis in the proof shows that even if $\alpha = 1$ and there is goal congruence, the principal may not prefer a better E if this has a sufficiently adverse effect on agent incentives (condition (ii) in the Proposition fails).

To see this more clearly, using our earlier functional form for g and assuming that $\gamma = 1$ and $b = 1$, we have:

$$u_P = EB + (1-E)(1-E)\alpha B$$

$$u_P = (1-E\beta)\alpha B + (1-(1-E\beta))EB$$

for the cases with and without AI authority. Then, the marginal benefit of increasing E is $(1-2(1-E)\alpha)B$ (AI authority) or $(2E-\alpha)B$ (H authority). Thus, even under AI authority, P does not always prefer a higher E. Indeed, for $E < 1 - \frac{1}{2\alpha}$, P would prefer a lower E and even for E up to $2 - \frac{1}{\alpha}$ may prefer not to employ an AI at all. Under H authority, it is only when $E \geq \alpha$ that the principal would employ an AI.

Intuitively, while it is the case that P would employ a perfect AI (with $E = 1$) and give it decision authority if that AI were available, when AI is imperfect, P may prefer to reduce the reliability of the AI as a means of encouraging more H effort. Note that the benefit to AI over H authority is $(1-\alpha)(1-2E)B$ which is decreasing in E. Thus, the lower is the performance of the AI (because of technical feasiblity or choice), the more likely is P to choose H rather than AI authority.

Another way to consider AI performance is from the perspective of bias. Suppose that even if the AI learns project payoffs, it does so imperfectly so that with probability μ it recommends P's preferred project but otherwise recommends a project with payoffs of $(\bar{\alpha}B, \bar{\beta}b)$ where $\bar{\alpha} \leq \alpha$ and $\bar{\beta} \leq \beta$. In this case, P and H payoffs under AI authority are:

$$u_P = E(\mu + (1-\mu)\bar{\alpha})B + (1-E)e\alpha B$$

$$u_H = E(\mu\beta + (1-\mu)\bar{\beta})b + (1-E)eb - g(e)$$

and under H authority are:

$$u_P = e\alpha B + (1-e)E(\mu + (1-\mu)\bar{\alpha})B$$

$$u_H = eb + (1-e)E(\mu\beta + (1-\mu)\bar{\beta})b - g(e)$$

Note that while H's effort does not change with μ under AI authority, under H authority it falls with μ. Intuitively, as more bias $(1-\mu)$ is introduced, the human agent is more motivated to avoid the AI making decisions as those decisions are more likely to be poor outcomes for H. A lower μ creates an AI that antagonizes H. Thus, even though a biased AI may not be preferred, ceteris paribus, by P, it may be employed under H authority so that the human agent relies less on the AI so long as $\frac{\beta-\bar{\beta}}{1-\bar{\alpha}}$ (i.e., the degree to which the AI choice harms H more than P) is sufficiently high.[5]

4 A Taxonomy

The potential trade-off that arises in terms of agent incentives when a better AI is employed, allows for a taxonomy of the types of AI that a principal might choose to employ. This taxonomy is show in Table 1 under the regimes of AI and H authority and whether P has a preference for better AI under each.

The different types of AI are as follows:

- **Replacement AI**: If a high performing AI is available (i.e., E close to 1 and sufficiently

[5] Of course, decreasing μ has to be considered to be a better option than switching back to AI authority. Using our earlier functional form, under H authority, $\hat{e}_H = (1 - E(\mu\beta + (1-\mu)\bar{\beta})b$ and, examining, $\frac{du_P}{d\mu}$ as $\mu \to 1$, we can see that it will be worthwhile to introduce bias if $(\beta - \bar{\beta})b(\alpha - E) > (1-(1-E)b)(1-\bar{\alpha})\alpha$. If this condition holds, then it is optimal to introduce some bias and employ an antagonistic AI if H authority is otherwise optimal with $\mu = 1$.

Table 1: AI Taxonomy

	AI Authority	H Authority
Better AI	Replacement AI	Augmentation AI
Worse AI	Unreliable AI	Antagonistic AI

unbiased), then the AI should hold decision rights and AI training focuses on eventually fully replacing humans.

- **Augmentation AI**: If current AI performance is relatively weak (E sufficiently low), human agents sufficiently well aligned with the principal, and human effort only weakly responsive to changes in AI performance, then human agents retain decision rights, and marginal improvements in AI performance or decreases in AI bias are profit-enhancing.

- **Unreliable AI**: When human agents are poorly aligned with the principal and potential AI performance is relatively strong, the AI optimally holds final decision rights. However, human effort is still important when the AI does not learn the optimal action, so if human effort is highly responsive to incentives, "unreliable" AI (lower E than technically feasible) is optimal as it trades off worse performance when the AI thinks it learns the optimal action against more human effort when it does not.

- **Antagonistic AI**: If current AI performance is relatively weak and human agents sufficiently well-aligned with the principal, but human effort strongly responds to changes in AI performance, then humans should retain decision rights. However, unlike with Augmentation AI, it is optimal to bias an AI such that the AI action is particularly bad for the agent. When the AI's choice "antagonizes" human agents, they increase effort to avoid the AI's recommendation being reported to the principal.

This taxonomy leaves many potential details out, but it maps the broad choices for organizations in terms of whether to give an AI or a human decision authority and, in turn, whether to favor a technically superior (i.e., reliable and unbiased) AI or not. This choice will depend on the nature of human reactions to working with AI as well as what is technically available to the organization.

On the latter point, we note here that the data that is used to train the AI may be relevant. For instance, replacement AI may require a high degree of reliability and, therefore, may require training based on repeated experiments rather than data that may be at hand. The same is true for augmenting AI although the organization may be more tolerant of data that is generated by past human decision observations. For unreliable AI, there may be reasons to forgo extensive data training while for antagonistic AI, data that identifies outcomes that humans dislike may be valuable. In future work, we will explore the issues of training data – in particular, how these interact with human incentives both past and present – in order to develop a clearer picture of the types of AI that may be employed at different stages of AI adoption in organizations.

References

Philippe Aghion and Jean Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997.

Philippe Aghion, Benjamin F Jones, and Charles I Jones. Artificial intelligence and economic growth. In Joshua Gans Ajay Agrawal and Avi Goldfarb, editors, *Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2019.

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

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 46:1–6, 2019.

Mustafa Dogan and Pinar Yildirim. Managing automation in teams. *Working Paper*, 2017.

Mustafa Dogan, Alexandre Jacquillat, and Pinar Yildirim. Strategic automation and decision-making authority. *Working Paper*, 2018.

Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and ai alignment. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.