AEA Papers and Proceedings 2020, 110: 80–84
https://doi.org/10.1257/pandp.20201034

ALGORITHMIC FAIRNESS AND BIAS[‡]

The Allocation of Decision Authority to Human and Artificial Intelligence[†]

By Susan C. Athey, Kevin A. Bryan, and Joshua S. Gans*

Artificial intelligence (AI) adoption is often equated with automation, with machines replacing humans in tasks and decisions. In practice, however, AI often augments 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. Prior knowledge can be important for anticipating outcomes in novel or unusual circumstances. In these contexts, the automated predictions of fully automated AI can be insufficient even when AI reduces the cost of prediction along some margins (Agrawal, Gans, and Goldfarb 2019). This motivates an analysis of precisely how humans and AIs would work together.

Using Aghion and Tirole (1997) altered to focus on human–AI interactions, we consider a principal who decides 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,[1] this is the first paper that focuses explicitly on the interaction of control problems for humans versus AI.

I. Model Setup

The initial model setup follows Aghion and Tirole (1997), where there is a principal (P) who allocates 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 can learn the agent's payoff for *all* projects with probability e. We assume disutility of effort is increasing and 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.

[‡]*Discussants:* Joshua Gans, University of Toronto; Avi Goldfarb, University of Toronto; Jorge Guzmán, Columbia University; Shane Greenstein, Harvard Business School.

*Athey: Stanford University and NBER (email: athey@stanford.edu); Bryan: University of Toronto, (email: kevin.bryan@rotman.utoronto.ca); Gans: University of Toronto and NBER (email: joshua.gans@utoronto.ca). Thanks to Jorge Guzmán for an excellent discussion.

[†]Go to https://doi.org/10.1257/pandp.20201034 to visit the article page for additional materials and author disclosure statement(s).

[1] See, for instance, Agrawal, Gans, and Goldfarb (2019) and Dogan, Jacquillat, and Yildirim (2018).

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.[2] Assume E to be common knowledge.

We consider the following timing. First, decision rights are allocated: either the AI or the agent is delegated formal final decision authority. Second, the agent chooses how much effort to exert. Third, the nondelegated player 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 zero suffices.

II. Allocating Decision Authority

Principal P's choice regarding whether to give H or the AI decision authority depends on their payoff in anticipation of H's choice of effort in learning about project payoffs. If H 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 their own (human) authority and the AI's authority respectively. These are determined by the following first-order conditions:

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

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

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 $(1-E)/\left(1-E\frac{1}{\alpha}\right) \geq \hat{e}_H/\hat{e}_{AI}$. As the right-hand side exceeds one, AI authority is optimal only 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. 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 nonnegative returns for P.

III. 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

[2] In Aghion and Tirole (1997), P is assumed to have the ability to learn project values with probability E provided they incur an effort cost. Here, we have endowed P with an AI that can learn on their behalf, and H knows the AI's capabilities.

to P as it reduces the 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 nondecreasing 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*) $\left|g'(\hat{e}_{AI})/\left(g''(\hat{e}_{AI})(1-\hat{e}_{AI})\right)\right| \leq 1$.

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, it is

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

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

To see this more clearly, assuming that $g(e) = (1/2)e^2$ and $b = 1$, we have $u_P = EB + (1-E)(1-E)\alpha B$ and $u_P = (1-E\beta)\alpha B + \left(1-(1-E\beta)\right)EB$ for the cases with and without AI authority. Then, the marginal benefit of increasing E is $\left(1-2(1-E)\alpha\right)B$ (AI authority) or $(2E-\alpha)B$ (H authority). Thus, even under AI authority, P does not always prefer a higher E. Indeed, for $E < 1 - \left(1/(2\alpha)\right)$, P would prefer a lower E, and even for E up to $2 - (1/\alpha)$, P 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 feasibility 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 it 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

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

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

and under H authority, they are

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

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

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 $(\beta - \bar{\beta})/(1 - \bar{\alpha})$ (i.e., the degree to which the AI choice harms *H* more than *P*) is sufficiently high.[3]

IV. A Taxonomy

The trade-off of human effort and decision alignment from decision rights and AI quality generates a taxonomy of optimal AIs. This taxonomy is shown in Table 1 under the regimes of AI and human authority, along with whether the principal 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 one and sufficiently 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 are sufficiently well aligned with the principal, and human effort is 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 are 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.

Table 1—AI Taxonomy

	AI authority	Human authority
Better AI	Replacement AI	Augmentation AI
Worse AI	Unreliable AI	Antagonistic AI

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 the AI as well as what is technically available to the organization.

On the latter point, we note here that the data that are 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 augmentation AI, although the organization may be more tolerant of data that are generated by past human decision observations. For unreliable AI, there may be reasons to forgo extensive data training, while for antagonistic AI, data that identify 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—to develop a clearer picture of the types of AI that may be employed at different stages of AI adoption in organizations.

[3] 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 $du_P/d\mu$ 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 $\mu = 1$.

REFERENCES

Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy* 105 (1): 1–29.

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

Dogan, Mustafa, Alexandre Jacquillat, and Pinar Yildirim. 2018. "Strategic Automation and Decision-Making Authority." https://dx.doi.org/10.2139/ssrn.3226222.