To Err is Algorithm: Algorithmic fallibility and economic organisation

Juan Mateos-Garcia

Nesta, 58 Victoria Embankment
London, EC4Y 0DS
Juan.mateos-garcia@nesta.org.uk

17 August 2017

Abstract

Algorithmic decision-making systems based on artificial intelligence and machine learning are enabling unprecedented levels of personalisation, recommendation and matching. Unfortunately, these systems are fallible, and their failures have costs. I develop a formal model of algorithmic decision-making and its supervision to explore the trade-offs between more (algorithm-facilitated) beneficial decisions and more (algorithm-caused) costly errors. The model highlights the importance of algorithm accuracy and human supervision in high-stakes environments where the costs of error are high, and shows how decreasing returns to scale in algorithmic accuracy, increasing incentives to 'game' popular algorithms, and cost inflation in human supervision might constrain optimal levels of algorithmic decision-making.

Keywords— algorithmic decision-making; data-driven decision-making; organisational design.

1 Introduction

Algorithmic decision-making systems, including those based on machine learning and Artificial Intelligence (AI), are rapidly being adopted across the economy and society, where they process fast-growing datasets to deliver personalised, interactive, 'smart' goods and services. Early excitement about the benefits of these systems have started to be tempered with concerns about their risks, particularly as they are adopted in highly sensitive domains like health, policing or education [1]. These concerns include algorithmic fairness when biases in the input data generate discriminatory decisions [2], potential manipulation of users [3] and 'filter bubbles' [4, 5].

But even in the absence of biases in their inputs, ill-intent on their designers, or emergent, hard to predict effects, algorithms make errors, and these errors have costs. This algorithmic fallibility is behind several recent controversies in online platforms:

- **YouTube advertising controversy**: YouTube placed adverts from global brands on videos with hate speech and offensive content. [6]
- **Facebook video controversy**: Facebook did not prevent the diffusion of violent content in its platform. [7]
- **Google autocomplete controversy**: The search autocomplete feature directed users looking for information about the Holocaust to far-right and Nazi websites [8]

This points to an important trade-off between more (algorithm-facilitated) beneficial decisions and more (algorithm-caused) costly errors, that should be taken into account by their designers, users and regulators.

Evidence to manage this societal trade-off is however lacking, and this explains recent calls for a programme of research on the impact of algorithmic decision-making *"a practical and broadly applicable social-systems analysis [that] thinks through all the possible effects of AI systems on all parties [drawing on] philosophy, law, sociology, anthropology and science-and-technology studies, among other disciplines"* [11]. Although Economics does not appear in this list, its emphasis on formalising the analysis of trade-offs has much to contribute to these important debates.

In this paper, I pursue this agenda by developing a economic model of algorithmic decision-making which builds on past analyses of economic organisations as information-processing systems composed of unreliable agents, and particularly on Sah and Stiglitz's work about the design of organisational architectures to manage the risks of this unreliability. [13, 12, 14]

After reviewing this literature in Section 2, I describe the model in Section 2, using it to explore the following questions:

- In what scenarios should we leave decisions to algorithms, and how accurate do those algorithms need to be?
- How can we organise humans and algorithms to achieve desired outcomes?
- What is the optimal number of decisions that these decision-making systems can make, and what factors determine their value?

Section 4 sets out some of the model's policy and organisational implications, and Section 5 concludes with a discussion of next steps.

2 Related Work

2.1 Economic agents as information processors

There is nothing new about the idea that economic agents and the systems they form act as computers that receive, process and communicate information. Hayek, for example, argued that the market system is superior to socialism because it uses information from prices to coordinate decentralised economic decisions, harnessing local pools of knowledge hard to access for decision-makers at the centre [15]. Traditional models of firm behaviour and markets (including the theory of general of equilibrium) describe the economic problem as a computation of the solution in a optimisation task [16].

This optimisation is constrained by prices that reflect material scarcity of natural resources, labour and capital, but not by information (which is assumed to be freely accessible), or by agent' ability of agents to compute that information (which is assumed to be unlimited). Perfect rationality and infallibility in decision-making follow on from these assumptions [17].

Herbert Simon challenged this view, contending that economic actors have limited cognitive and computational capabilities, and that this makes their rationality 'bounded' rather than perfect [18]. This means that cognitive resources are scarce and need to be allocated carefully [19]. One way in which organisations do this is by developing routines and heuristics to automate behaviour. This creates an important economic function for algorithms that encode those routines, allowing humans to focus on those activities where they have a comparative advantage.

This analysis also implies that organisations with better architectures to process information should outperform others. These architectures can be modelled as networks whose nodes are workers and managers that process information and make decisions. [20, 21].

2.2 Human fallibility and economic organisation

In several papers published over the 1980s, Sah and Stiglitz adopted this approach to compare the economic performance of two organisational architectures: *hierarchies* (where decision-making requires approval by different layers in the organisation), and *polyarchies* (where any one actor in the organisation can approve a decision) [13, 12].[1]

Their analysis showed that hierarchies select fewer 'good' projects, and polyarchies tend to select more 'bad' projects. Inventors who generate new projects in a hierarchy will become more risk-averse, potentially making that economic system less innovative.

Recent expansions of Sah and Stiglitz's analysis have focused on the design of complex organisational structures formed by mixes of hierarchies and polyarchies with the aim of reducing error [22, 23, 24]. Interestingly, although these studies explicitly link organisational design problems with the engineering of reliable technical systems from unreliable components [25, 26], they do not consider the possibility that organisational systems might be formed of both human and non-human components, as I do here.

2.3 The economics of data-driven decisions and predictions

The 'big data explosion' has brought with it substantial interest on the economic impact of data, and on the organisational structures which complement (algorithmically-enabled) 'data-driven decision-making'. Past studies have shown a positive relationship between company performance and data-driven decision-making, which is also strongly complementary to decentralised decision-making and employee empowerment [9, 10]. One explanation for

---

[1] In doing this, they were assessing the relative merits of socialist economic systems (hierarchies), and market-based systems (polyarchies).

this is that in data-driven organisations, employees can access organisation-wide knowledge to inform their decisions without having to consult with others (including managers). However, these studies do not consider in detail how to organise systems of algorithms and humans for decision-making.

A recent exception to this is Agrawal et al's analysis of the economic impact of AI [27]. There, they define AI as a technology that increases the supply of predictions, and the number of risky decisions that an organisation can make. They consider the complementarity between AI and human judgements about the probability of different pay-offs from decisions, showing that human judgement can be particularly valuable when it detects negative pay-offs for a risky algorithmic decision, helping to avoid 'bad' scenarios.

3 The Model

3.1 A single agent makes a decision

We consider an organisation that needs to process informational inputs, making decisions about their quality. Imagine for example a list of transactions where some of them are fraudulent, or a list of news items where some of them are 'fake'. To begin with, there is a single algorithm a_1 processing this information.[2] The algorithm receives an input and decides whether to accept it or not. The input can be 'good' or 'bad', and the proportion of good inputs in the initial pool of inputs is α.

The algorithm accepts good inputs with probability p_{11} (true positive rate) and bad inputs with probability p_{12} (false positive rate). If accepted, good inputs generate a benefit of r_1 and bad inputs generate a negative cost of $-r_2$.

The expected value of a decision d_1 is:

$$\mathbf{E}(d_1) = \alpha p_{11} r_1 - (1-\alpha) p_{12} r_2 \tag{1}$$

This value is positive if:

$$\frac{\alpha p_{11}}{(1-\alpha) p_{12}} > \frac{r_2}{r_1} \tag{2}$$

This represents a trade off between the risks of an algorithmic decision and its expected payoff. It is positive if the odds of accepting a good input relative to accepting a bad input outweigh the costs of bad inputs relative to the benefits of good inputs. Higher quality inputs and better algorithmic accuracy, and declines in the cost to benefit ratio improve the expected value of the decision and the algorithm that enables it. An implication is that algorithms making decisions in high stake environments (where the cost of an error is large) should have high accuracy. [3]

3.2 Supervision and filtering

Now we consider the situation where a human supervisor a_2 enters the picture. The supervisor takes the accepted inputs from the algorithm with probability t (this represents the frequency of supervision), and processes them with a true positive rate p_{21} and a false positive rate p_{22}. From one point of view, a_2 is supervising the decisions of a_1. From another, a_2 is filtering inputs for a_2.[4]

The expected benefit of a decision d_2 is:

$$\begin{aligned}\mathbf{E}(d_2) = \alpha r_1 (p_{11} t p_{21} + (1-t) p_{11}) - \\ (1-\alpha) r_2 (p_{12} t p_{22} + (1-t) p_{12})\end{aligned} \tag{4}$$

The net contribution of the human supervisor is positive when the improvement in expected value for d_2 outweighs the increase in costs. When we compare this with her cost per decision $C_2(t)$ we get:

$$(1-\alpha) r_2 t p_{12} (1-p_{22}) - \alpha r_1 t p_{11} (1-p_{21}) > C_2(t) \tag{5}$$

The value of supervision is the benefit of rejecting a bad input erroneously accepted by the algorithm minus the cost of rejecting a good input accepted by the algorithm. This value increases when the pool of projects worsens, when the algorithm loses accuracy, and when the costs of bad inputs increases relative to the rewards of good inputs.

3.3 Scale

Organisational profits are:

$$\Pi(n) = n\mathbf{E}(d_2) - C(n) \tag{6}$$

---

[2] I use the term 'algorithm' to refer to technologies that turn informational inputs into predictions (and depending on the system receiving the predictions, decisions). There are many processes to do this, including rule-based systems, statistical systems, machine learning systems and Artificial Intelligence (AI).

[3] This set-up represents a binary classification problem as in [13] where an organisation processes items of two types. If the organisation was making recommendations or match-making, we could consider a measure of overall accuracy p instead a false positive or false negative rate. Equation 2 would be:

$$\frac{\alpha p}{(1-\alpha)(1-p)} > \frac{r_2}{r_1} \tag{3}$$

Here, the odds of making a correct decision have to be higher than the costs of making the wrong decision.

[4] This set-up could be modified to represent a situation where the supervisor verifies the inputs rejected by an algorithm. We could then compare the expected value of this design with one where the supervisor checks the 'good' outputs from the algorithm.

How do they change with n?

When looking at this question, we make the following assumptions: r_1, r_2 and t do not change with the number of inputs, and the human supervisor is a perfect screen ($p_{21} = 1$ and $p_{22} = 0$). We also assume that changes in the accuracy of the algorithm are symmetric: a_1 $\delta p/\delta n$ = $\delta p_{11}/\delta n$ = $-\delta p_{12}/\delta n$. The true positive rate changes in the opposite direction of the false positive rate, and at the same rate, as n grows.

Taking all of this into account:

$$\begin{aligned}\frac{\delta \mathbf{E}(d_2)}{\delta(n)} = \mathbf{E}(d_2) + n(\frac{\delta p}{\delta n}(\alpha r_1 + (1-\alpha)(1-t)r_2)+ \\ \frac{\delta \alpha}{\delta n}(r_1 p_{11} + r_2 p_{12}(1-t)))\end{aligned} \tag{7}$$

If the algorithm's accuracy and the quality of the input pool α improve with the number of inputs, then supervision makes a smaller contribution to profits because the number of false positives checked and fixed by the supervisor declines. The opposite is true if the accuracy of the algorithm and/or the quality of the input pool α decline with n.

Which set of assumptions is more likely? On the one hand, new machine learning techniques help algorithms benefit from large training datasets. On the other, more decisions can also degrade an algorithm's accuracy, for example if they force it to make predictions about unusual and novel inputs. Further, when an algorithm becomes very popular (makes more decisions), users might have more reasons to try to 'game it', bringing a decline in α.

All this means that returns to scale in algorithmic decision-making depends on context and change over time in ways that could be endogenous to the actions of algorithm designers. Regarding $\delta C/\delta n$:

$$\frac{\delta C}{\delta n} = \frac{\delta C}{\delta L_{a1}}\frac{\delta L_{a1}}{\delta n} + \frac{\delta C}{\delta L_{a2}}\frac{\delta L_{a2}}{\delta n} \tag{8}$$

Here, the number of inputs verified by the supervisor, n_2, depend on the total number of inputs processed by the algorithm, supervision frequency, the quality of the input pool and algorithmic accuracy:

$$\begin{aligned}\frac{\delta n_2}{\delta n} = t\big[(\alpha p_1 + (1-\alpha)p_2)+ \\ n(\frac{\delta \alpha}{\delta n}(p_{11} - p_{12}) + \frac{\delta p}{\delta n}(2\alpha - 1))\big]\end{aligned} \tag{9}$$

We refer to it as $n'_2(p', \alpha')$.

w_1 is the salary of the algorithm developers, and w_2 is the salary of the supervisors. Their respective marginal productivities are z_1 and z_2.

$$\frac{\delta C}{\delta n} = \frac{w_1}{z_1} + \frac{w_2}{z_2} n'_2(p', \alpha') \tag{10}$$

Costs depend on the salaries and marginal productivities of developers and supervisors, and the number of inputs a_1 which the supervisors have to verify. Increases in the quality of the input pool increases this number, and increases in the accuracy of the algorithm will increase the workload for a_2 if $\alpha > 0.5$ (in this case, the number of true positives she has to check grows faster than the decline in true negatives).

Although algorithm developers are likely to be get paid more than supervisors, their marginal productivity will correspondingly be higher: a single algorithm, or an improvement in an algorithm, can be scaled up over millions of inputs. By contrast, supervisors need to check each input individually. This means that as the number of inputs grows, supervisor costs could be expected to gain relative importance, potentially limiting the efficient scale of algorithmic decision-making.

4 Implications

The model has highlighted some important factors and dynamics in algorithmic decision-making situations. I overview some of their implications here:

4.1 Algorithmic decision-making in environments with different stakes

First, I have shown that algorithms making decisions in situations where the stakes are high need to be very accurate to avoid costly failures. By contrast, inaccurate algorithms might be suitable for situations where the costs of error are low, or where the quality of the input pool is high.[5]

This also means that organisations in 'low-stakes' environments can experiment with novel algorithms, including some that begin with low-accuracy. As these are improved, they can be transferred to 'high stake domains'.[6] Algorithms need to be adopted more carefully in domains where the costs of errors are high, such as health or the criminal justice system, and when dealing with groups who are more vulnerable to algorithmic errors.[1] Only highly accurate algorithms will be suitable for these risky decisions, unless they are complemented with expensive human supervision to detect and fix errors. This could create natural limits to algorithmic decision-making in these domains.

---

[5] For example, the recommendation engines in online platforms like Amazon or Netflix often make irrelevant recommendations, but the cost of those errors is relatively low – they tend to be simply ignored.

[6] The technology companies that develop these algorithms often release them as open source software for others to download and improve, making these spill-overs possible.

If policymakers want to remove these barriers, they should invest on R&D to improve algorithmic accuracy, encourage the adoption of high-performing algorithms from other sectors, and experiment with new organisational designs to identify and remove errors.

Even commercial organisations are not immune to some of these problems: For example, YouTube has started blocking adverts in videos with less than ten thousand views in response to the controversy highlighted in the Introduction. In those videos, the benefits of correct algorithmic ad-matching are low because of their limited audiences, and the average quality of inputs is low, which creates the risk of posting adverts against offensive content. Meanwhile, Facebook recently announced that it is hiring 3,000 human supervisors (almost a fifth of its workforce as of 2016) to moderate content in its platform.

4.2 The costs and benefits of keeping humans in the loop

The model shows that human supervision of algorithms can be costly. The marginal productivity of human supervisors is likely to be low compared to that of algorithm developers, potentially constraining the optimal scale for algorithmic decision-making (as long as we hold supervision frequency constant).[7]

One potential strategy to contain these costs is to outsource supervision to users, for example by providing them with tools to report low quality inputs. YouTube, Facebook and Google have all implemented these tools in response to their algorithmic controversies. However, this is not a risk-free strategy: delegating content supervisions to users can be highly detrimental to their experience inside the platform.

Human supervision remains valuable even when the risks from algorithmic error are low. The reason for this is that it provides a buffer against sudden declines in organisational performance if the accuracy of algorithms decreases or the quality of inputs suffer a negative shock. When this happens, the number of erroneous decisions detected by humans and the benefit of fixing them increase. Human supervisors can also call attention to these 'algorithmic crises', potentially helping to identify and address problems more rapidly.

This 'early warning sign' or homeostatic feature of human supervision can be particularly important if algorithmic errors create costs with a delay, or costs that are hard to measure (for example if erroneous recommendations result in self-fulfilling prophecies, or costs that are incurred outside the organisation).[8]

5 Conclusion

5.1 Extending the model

The model could be extended to capture other scenarios:

For example, I have assumed that there is certainty about benefits and costs, and that these are constant for different decisions. We could instead model them as a random variable. This would allow us to incorporate algorithmic fairness and bias into the analysis - If algorithmic errors are more likely to affect vulnerable groups (who will suffer higher costs), and these errors are less likely to be detected, this could increase the expected cost of errors, requiring more accurate algorithms or more human supervision in these circumstances.

I also assumed that the benefits (r_1) and costs (r_2) per decisions do not change as the number of inputs increase. One could however imagine scenarios where benefits increase with n (e.g. if an organisation gains market power), or where the costs of errors increase (e.g. if the organisation and the errors it makes become more visible). Some of these scenarios might have important implications for our understanding of competitive (and anti-competitive) dynamics in data-driven markets.

The behaviour of agents who influence the quality of the input pool is currently exogenous to the model. We could bring these bahaviours and the incentives that shape them into the model, and use it to explore how the interactions between those agents (many of which are also algorithmic!) and algorithmic decision-makers shape the supply of low quality informational inputs (spam, scams, fake news, illicit content etc.) in digital markets.

5.2 Operationalisation

When reviewing the literature, I mentioned previous extensions of Stiglitz and Sah's work that explored how different organisational achitectures can be integrated to reduce error. The model of algorithmic decision-making I have developed here could be expanded in a similar way.

---

[7] This situation has similarities with the 'unbalanced growth' model used in the analysis of economies composed of fast productivity growth and low productivity growth sectors [29]

[8] For example, in the YouTube advertising controversy, the significant reputational cost from previous errors was only incurred when brands noticed that their adverts had been posted against offensive content. The 'fake news controversy' after the US election is another example of hard to measure costs: algorithms' inability to discriminate between real news and hoaxes creates externalities, potentially justifying stronger regulations and more human supervision.

In addition to shedding light on some important factors and dynamics in algorithmic decision-making, an effort to operationalise the model and its extensions with data from specific domains or organisations could inform the design of better algorithmic decision-making systems, along the lines of [24]. There are many opportunities to integrate economic modelling with experimental and simulation methods as part of an effort to bring Economics into ongoing analyses of the impact of algorithms to which, as I have suggested in this paper, the discipline has much to contribute.

Acknowledgements

This paper benefited from comments from Geoff Mulgan, and was inspired by conversations with John Davies at Nesta.

References

[1] O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Books.

[2] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.

[3] Flick, C. (2016). Informed consent and the Facebook emotional manipulation study. *Research Ethics*, 12(1), 14-28.

[4] Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

[5] Alstyne, V. M., & Brynjolfsson, E. (1997). Electronic Communities: Global Villages or Cyber Balkanization?. In *Proceedings of the Seventeenth International Conference on Information Systems* (pp. 373-379).

[6] Solon, O. (2017). Google's bad week: YouTube loses millions as advertising row reaches US. The Guardian, 25 March 2017.

[7] Solon, O. (2017). Live and death: Facebook sorely needs a reality check about video. The Guardian, 26 April 2017.

[8] Cadwalladr, C. (2016). Google, democracy and the truth about internet search. The Guardian, 4 December 2016.

[9] Brynjolfsson, Erik, Lorin M. Hitt, and Heekyung Hellen Kim. "Strength in numbers: How does data-driven decisionmaking affect firm performance?." (2011).

[10] Bakhshi, H., Bravo-Biosca, A., & Mateos-Garcia, J. (2014). The analytical firm: Estimating the effect of data and online analytics on firm performance. Nesta Working Paper 14/05

[11] Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538, 311-313.

[12] Sah, R. K., & Stiglitz, J. E. (1985). Human fallibility and economic organization. The American Economic Review, 75(2), 292-297.

[13] Sah, R., & Stiglitz, J. E. (1984). The architecture of economic systems: hierarchies and polyarchies. NBER Working Paper No. 1334

[14] Sah, R. and Stiglitz, J. (1988). Committees, hierarchies and polyarchies. Economic Journal, 391 (8), 451-470.

[15] Hayek, F. A. (1945). The use of knowledge in society. The American economic review, 519-530.

[16] Varian, H. (1992). Microeconomic theory. W. W. Norton & Company, New York.

[17] Nelson, R. R., & Winter, S. G. (2009). An evolutionary theory of economic change. harvard university press.

[18] Simon, H. A., & Barnard, C. I. (1959). Administrative Behavior: A Study of Decision-making Processes in Administrative Organizaton. Macmillan.

[19] Simon, H. A. (1996). The sciences of the artificial. MIT press.

[20] Mount, Kenneth R., and Stanley Reiter. Computation and complexity in economic behavior and organization. Cambridge University Press, 2002.

[21] Radner, Roy. "The organization of decentralized information processing." Econometrica: Journal of the Econometric Society (1993): 1109-1146.

[22] Ioannides, Y. M. (1987). On the architecture of complex organizations. Economics Letters, 25(3), 201-206.

[23] Ioannides, Y. M. (2012). Complexity and organizational architecture. Mathematical Social Sciences, 64(2), 193-202.

[24] Christensen, M., & Knudsen, T. (2010). Design of decision-making organizations. Management Science, 56(1), 71-89.

[25] Moore, E. F., C. E. Shannon. 1956a. Reliable circuits using less reliable relays, part I. J. Franklin Inst. 262(September) 191–208. 89

[26] Moore, E. F., C. E. Shannon. 1956b. Reliable circuits using less reliable relays, part II. J. Franklin Inst. 262(October) 281–297.

[27] Agrawal, A., Gans, J. and Goldbarb, A. (2017). Exploring the Impact of Artificial Intelligence: Prediction versus Judgment. Working paper presented at the American Economic Society Annual Meeting, January 2017.

[28] Solon, O. (2017). Facebook killing video puts moderation policies under the microscope, again. The Guardian, 17 April 2017.

[29] Baumol, W. J., & Bowen, W. G. (1993). Performing arts-the economic dilemma: a study of problems common to theater, opera, music and dance. Gregg Revivals.