Interdependency in AI Decision Making: Challenges in Regulation

Ellie F. Baker and Munther Dahleh

In many cases, greater consideration for interdependency between components of an AI decision making system could improve regulation, implementation, and research. The aim of this paper is twofold; first, to show that consideration of interdependency is crucial to resolving problems involving AI. To do so, we review important research that considers AI interdependencies in order to identify and address current problems in AI decision making systems. Second, we aim to empower practitioners, researchers, and regulators to consider AI interdependencies in their work. To do so, we propose a full-cycle model of AI decision making systems, and map the problems and interdependencies discussed in our literature review onto our model.

Institute for Data Systems and Society, Massachusetts Institute of Technology

I. Introduction

The literature reviewed in our Article discover and address problems involving AI by studying interdependencies between components in an AI Decision Making System (hereafter AI System). For example, consider algorithmic monoculture; adoption of a particular algorithm may be the best response for a firm, acting independently (Kleinberg and Raghavan, 2021). But this choice, replicated by many or all firms in a market, has negative welfare consequences (ibid.). The problem is current regulatory systems - whether algorithmic, legal, market-based, or social (Lessig, 1998) – may not adequately consider AI interdependencies. For example, many audits of algorithmic systems assess outputs of algorithms but not outcomes of algorithm deployment. At minimum, auditing algorithmic deployment requires understanding how an algorithm interacts with people or firms to change their decisions, it could also involve understanding how algorithmic adoption changes the current equilibrium. Or consider an important critique of current antitrust law, that a singular focus on the price of consumer goods as indicators of anticompetitive behavior has resulted in a failure to address potentially anticompetitive conduct that exploits interdependencies for profit without raising consumer prices in the short term (Khan 2017).

There is a need and opportunity to update regulatory systems to consider AI interdependencies. In this paper we take a step towards that goal in two parts. First, we propose a full-cycle model of AI Decision Making with feedback and bring important System and Control Theory concepts to bear on our model. Our model builds on the AI model described in Kleinberg et al.'s "Discrimination in the Age of Algorithms" and the OECD AI System Model, both of which we discuss in Section II. Our model's specification of system components and their relationships may help practitioners consider AI system interdependencies moving forward and reveal the multiplicity of interventions that are possible to address current problems involving AI. Second, we survey the literature that identifies and addresses problems involving AI system interdependencies. We map the problems identified in this literature, and the interventions that would address them, to our model of AI Decision Making.

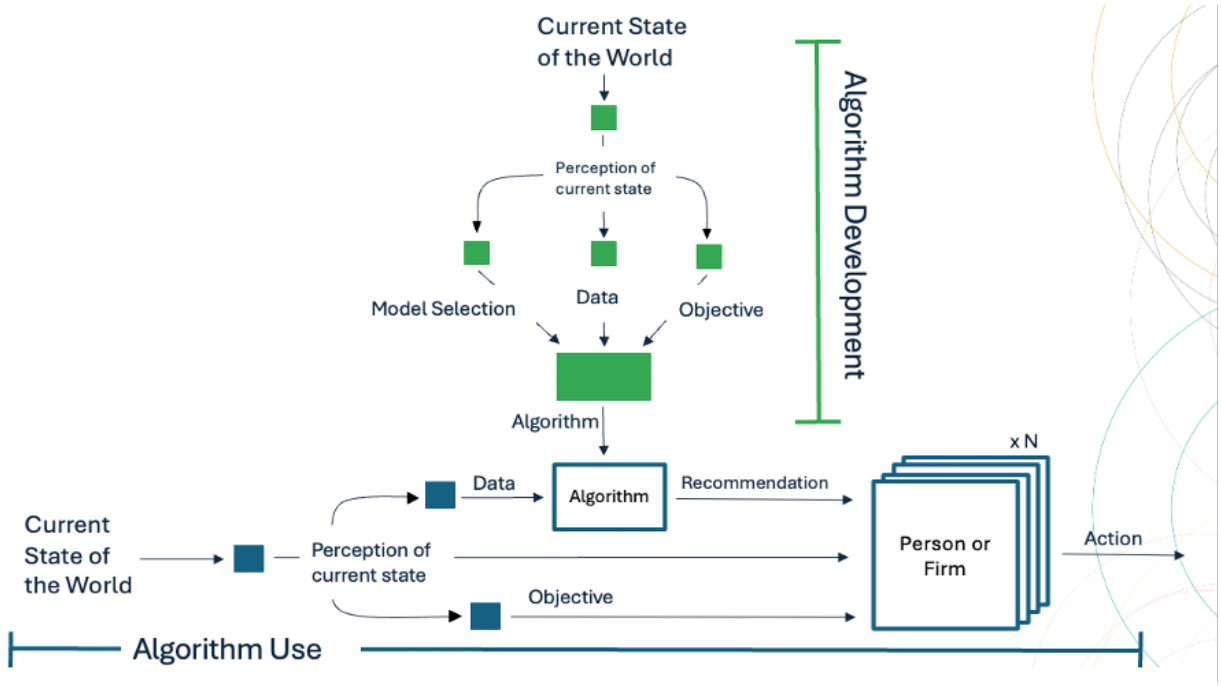

Figure 1.

II. A Model of AI Decision Making

Our full-cycle model of AI decision making (Figure 1) comprises two parts: algorithm development and algorithm use. We depict these on different axes to indicate that they may not occur concurrently and may be undertaken by different agents (agents are firms or individuals). In the diagram on the left, boxes represent 'operators' indicating that inputs "are transformed into" outputs through an unspecified process. Below we define components of our model that may not be readily apparent. The *objective* is a quantified goal to be optimized for (when used by an agent, the *objective* may be unconscious and imperfectly optimized). *Model selection* defines the selection of the architecture for the class of models from which to build a ML algorithm – this selection often entails making certain assumptions that may be ignored or missed at deployment. For example, one could use OLS regression to develop an algorithm that optimizes for a particular objective function based on input data, or one could use a neural network – the prediction function in the first case will be linear and have embedded structural assumptions, and in the second case it will have fewer embedded structural assumptions and is likely to be nonlinear. An agents' *perception of the world* is the information available to or inferred by an agent about the true state of the world. No agent has a complete understanding of the true state of the world, and many agents do not have the same perception of the true state of the world. The *current state of the world* represents all aspects of the unknown, true state of the world.

Our model builds on two existing models of AI Decision Making. The first is described in "Discrimination in the Age of Algorithms"; this model focuses on algorithm development and use - we extend this model to represent what may occur after an algorithmic recommendation is

produced and to incorporate feedback. The second model our work builds on is an AI system model published by the OECD. Our model involves a number of components also included in the OECD model. However, in contrast to the OECD model we include model selection in the algorithm development phase, we emphasize the role of human or firm decision making in determining actions that influence the current state of the world (including additional inputs to this decision other than an algorithmic recommendation). We also represent the system in a manner standard to control theory.

Control Theory for AI Decision Making Systems

Concepts from Control Theory formalize important AI decision making system dynamics. For example, consider feedback – a dynamic central to Control Theory in which outputs from the existing system (in this case, *actions*) shape what the system will look like in the future. A key challenge in Control Theory is ensuring feedback does not destabilize a system. However, in important cases, feedback compounds harm in a system. Cascading and correlated failures are two additional concepts from control theory with relevance to AI decision making systems. Discussed in the work we highlight on algorithmic monoculture, cascading failure occurs when failure in one component of a model produces failure in another component which in turn produces failure in another component and so on - resulting in a cascading effect. In turn, correlated failure occurs when multiple components of a model have similar structures, such that failure of a single type has widespread impact - systems which exhibit this characteristic have limited robustness.

III. Challenges and Interventions

We highlight challenges involving AI system interdependencies and the interventions that might address them until all components of our model have been included in at least one example. For each challenge, we highlight the components of our AI Decision-making system that are relevant to each of the examples we discuss. We do not distinguish between components that "cause" the challenges and components that "solve" the challenges we discuss. This is because any component that could be changed to "solve" the challenge can be considered a "cause" of the challenge (since it is not configured in a manner that solves the challenge at this moment). Making such a distinction might unintentionally limit thinking on the subject, since it can lead to a singular focus on one component of an AI system when many may be relevant.

Integrating experts and algorithms

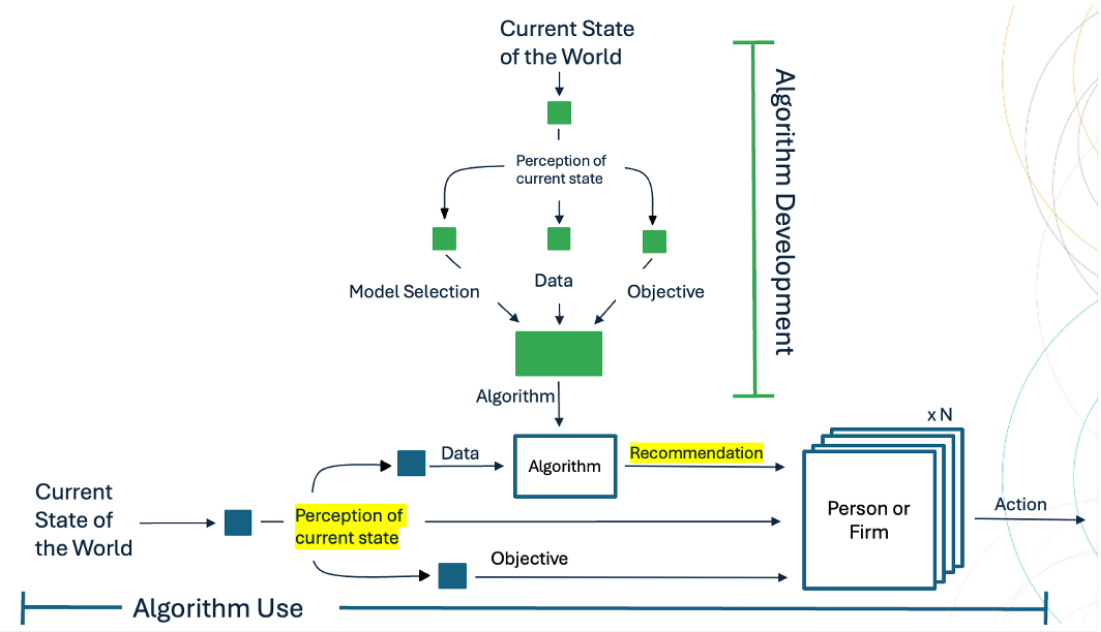

Figure 2.

In "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology" Agarwal et al. describe how in a controlled experiment, AI diagnoses from their ML model are more accurate than 75 percent of radiologists' diagnoses. Nonetheless, on average providing radiologists with AI predictions had no effect on their diagnostic accuracy (Agarwal et al., 2023). When AI predictions are relatively uncertain, they decrease radiologist decision quality – decision quality also declines when radiologists who are relatively certain about their diagnoses are provided with AI predictions (Agarwal et al., 2023). The authors provide strong evidence to indicate that this pattern is the result of correlation neglect (Enke and Zimmerman, 2019, in Agarwal et al., 2023; Agarwal et al., 2023). Correlation neglect occurs because radiologists' *perception of the current state of the world* is flawed – they act as though AI predictions, or *recommendations*, provide them with completely new information, when in reality much of the information provided by such predictions is a repackaged version of the information radiologists already use to make their decision (ibid.). Agarwal et al. point to the important implications correlation neglect has for optimal human-AI system design, including the fact that it is not always optimal to provide radiologists with AI recommendations. One intervention that could mitigate the problem the authors identify would be to provide ML *recommendations* only under certain conditions – for example when the ML prediction meets a specific certainty threshold (Agarwal et al., 2023).

Algorithmic monoculture

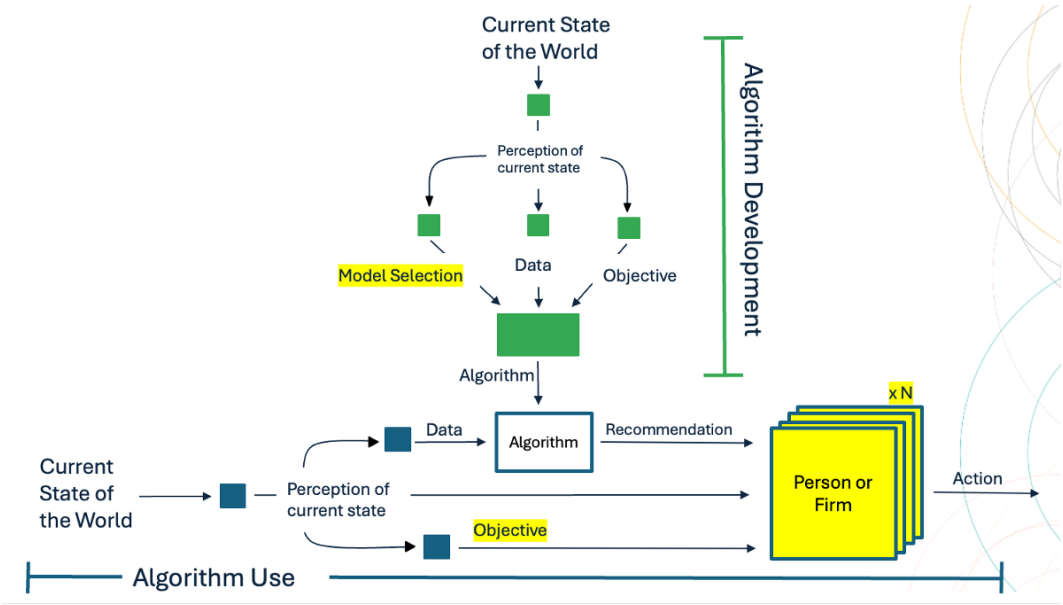

Figure 3.

Choices that may be optimal for an individual or firm acting independently and that are rationally adopted by many or all firms in a market may result in a suboptimal equilibrium. For example, mass adoption of the same algorithm can increase competition and the risk of correlated failure for firms and the chances of systematic exclusion for individuals (Kleinberg and Raghavan, 2021). Correlated failure is a term often associated with agriculture – if crops are not diverse enough, a single disease can wipe them all out (Kleinberg and Raghavan, 2021). Similarly, an economic environment may be more vulnerable when *many firms* adopt identical decision-making tools (ibid.). Correlated failures, while potentially catastrophic, are also relatively infrequent. Another consequence of algorithmic monoculture is constant - when firms adopt the same algorithms, their decisions are more similar, which can lead to increased competition (Kleinberg and Raghavan, 2021). For example, if five firms all adopt the same hiring algorithm, they will now all compete for the same first-choice candidate, when they might otherwise have had different top choices (Kleinberg and Raghavan, 2021). Homogenization of candidate rankings also harms some job applicants (ibid.). Individuals who are promising candidates under some, but not all, evaluation criteria may have fewer opportunities when all firms adopt a single evaluation criterion than in an algorithmically diverse market (Kleinberg and Raghavan, 2021).

One promising intervention to address these problems involves *model selection* – one could replace deterministic algorithms with stochastic algorithms that include "good noise" (for a comprehensive discussion and development of "bad noise" see Kahneman et al., 2022). A number of reinforcement learning algorithms could introduce such noise – these algorithms contain a degree of exploration (in which the algorithm searches for promising new options and selection from such options introduces "good noise") in addition to exploitation (in which the

algorithm prioritizes options that are already known to be successful). We discuss implementation of one such algorithm in a later section of this paper - in the implementation, Li et al. show that such an algorithm solves a number of other AI system challenges discussed in that section (Li et al., 2020). Exploration in our current context might mean selecting an individual whom the algorithm predicts to have high potential but on whom limited prior performance data exists due to past hiring decisions; for example, if a firm historically mainly hired students who graduated from Ivy League schools, this could be a student with excellent work experience and grades who attended a Community College (Li et al. 2020). Exploitation would be selecting a student with good grades who graduated from an Ivy League school (ibid.). Implementation of such algorithms would produce variation across firm's candidate rankings, thus reducing competition between firms for candidates and decreasing the likelihood of systematic exclusion and correlated failure.

Anticompetitive Behavior

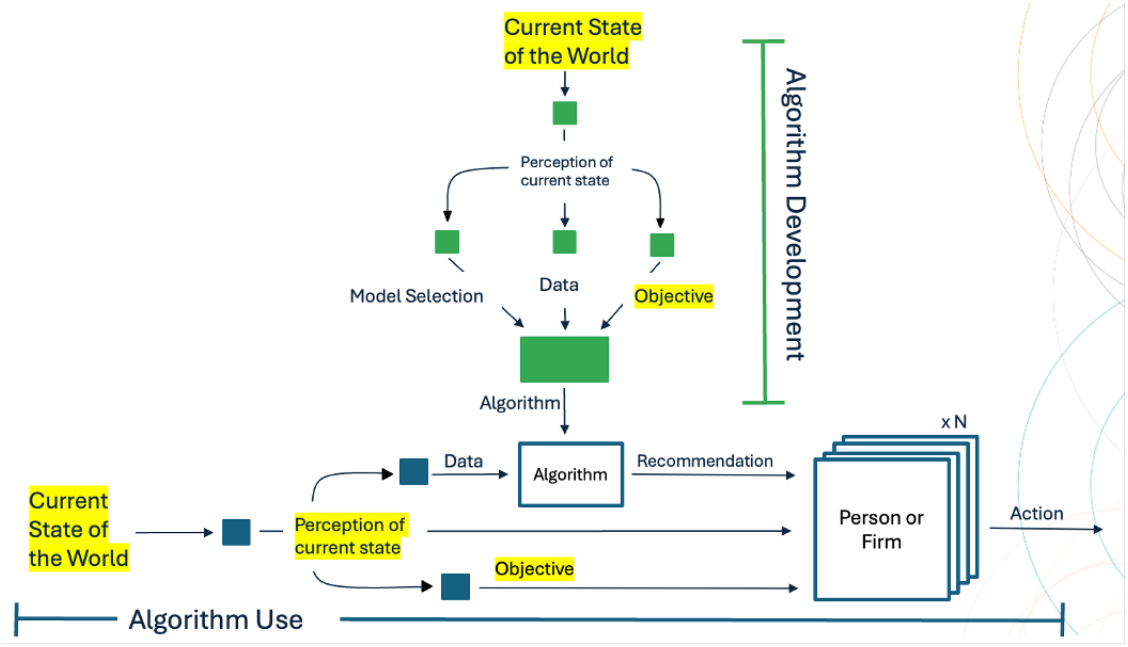

Figure 4.

Firms that control some aspect of the current state, say, a retail platform, and compete with other firms on that same platform, have profound power. Algorithms facilitate this power. For example, consider Amazon. Amazon operates and develops the Amazon retail platform and is also a retailer on this platform. We review three practices of Amazon, each facilitated by algorithms, that might be considered anticompetitive.

1. Amazon's ownership of multiple components of this system provides it with a much more expansive and accurate *perception of the current state* of the world, which means Amazon also has more accurate and expansive *data* than its competitors (Khan 2019).

One way Amazon has profited from this information gap is Amazon's adoption of similar features or product offerings of smaller companies on its platform that enjoyed early success – while it is likely such is likely a result of Amazon's expansive access to (ibid.).

2. In many cases, Amazon's configuration of its search results preferences Amazon products over other firms' products, holding other observable factors (like product ratings and delivery speed) constant (Farronato et al., 2023).
3. Project Nessie is an algorithm developed by Amazon to identify retailers who implement pricing algorithms that set their prices based on comparable Amazon goods (FTC, 2023). When Amazon identifies a critical mass of such retailers selling a certain product, they raise the price of their relevant product, which induces an across the board, sustained, price increase (FTC, 2023). Project Nessie and the actions it facilitated are estimated to have earned Amazon over $1 billion in profit (FTC, 2023).

In "The Separation of Platforms and Commerce", Lina Khan argues for increased emphasis on 'structural separatism' in antitrust. She advocates for policy and legal practice that scrutinize, and limit, firms' control over multiple components of AI Decision-making systems. This strategy would eliminate firms' incentives to design architecture that prioritizes their products (changing the objective function of such firms) (Khan 2019). Khan's intervention would also limit the possibility of dramatic information imbalance between competing firms and make the information imbalance that exists between a platform provider and firms that compete on the platform less harmful. Note that Khan's intervention does not directly prevent another firm from using an algorithm like Project Nessie to manipulate prices – however limiting market power reduces the chances that firms set prices based on a single 'price setter'. Developing additional statistical tools and policy to identify and address problematic algorithms like Project Nessie is an important issue for future research.

Algorithmic (mis)alignment

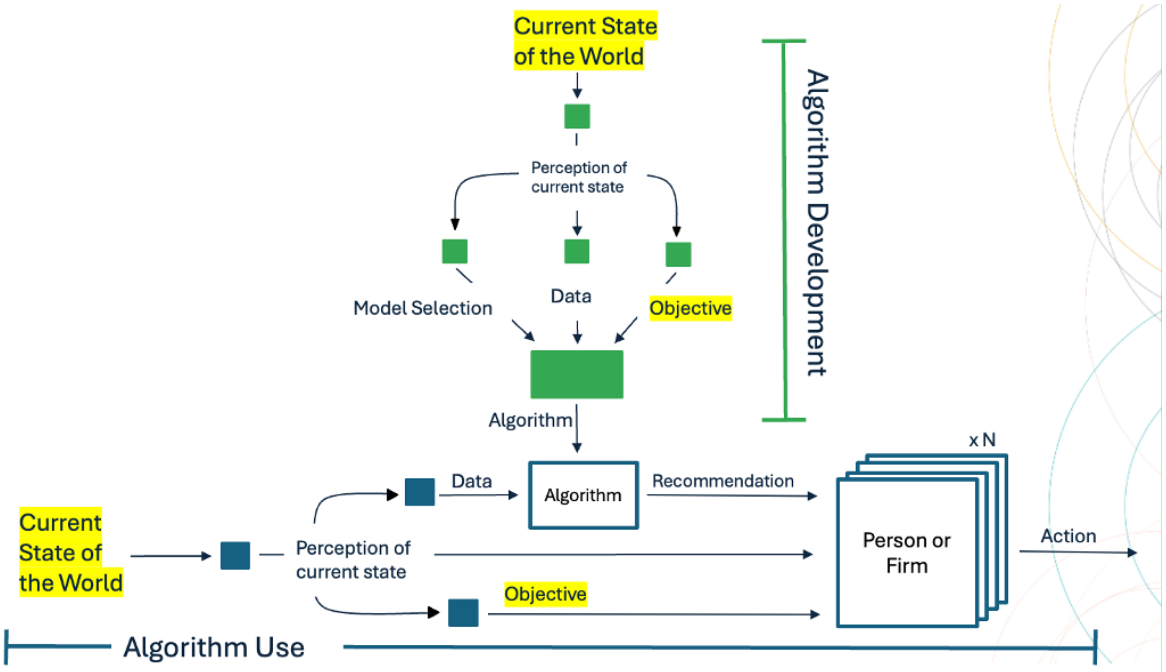

Figure 5.

Risk assessment algorithms are an important part of the US healthcare system (Obermeyer et al., 2019). They help dictate which individuals are selected for enrollment in risk-mitigation programs (ibid.), and they may also inform medical recommendations and determine health insurance premiums. Risk assessment algorithms are also biased (Obermeyer et al., 2019). The authors show that on average, if a Black person and a white person have the same risk score, the Black person will be more sick than the white person when the health of both people is assessed. This bias results in significant harm; for example, some Black people will be excluded from risk-mitigations programs for which they are eligible (ibid.). Changing the *objective* of commonly used risk scoring algorithms could significantly reduce the bias (Obermeyer et al., 2019). At present, risk scoring algorithms use future healthcare spending as a proxy for health – when health is the *objective* used by a person or firm to make patient or insurance decisions (Obermeyer et al., 2019). While healthcare spending and health are correlated with one another, they are not equivalent (ibid.). While both health and healthcare spending are functions of our *current state* with systemic racism, the authors show that a risk scoring algorithm that predicts health, rather than healthcare spending, is more accurate and significantly reduces racial bias in risk scoring algorithms (to understand how systemic racism influences healthcare spending, consider well-documented cases of racism in current medical practice that might reduce care and medical spending for Black patients, and the implication of the racial wealth gap for healthcare spending– which itself is caused by historical and current racism).

Designing more effective objective functions

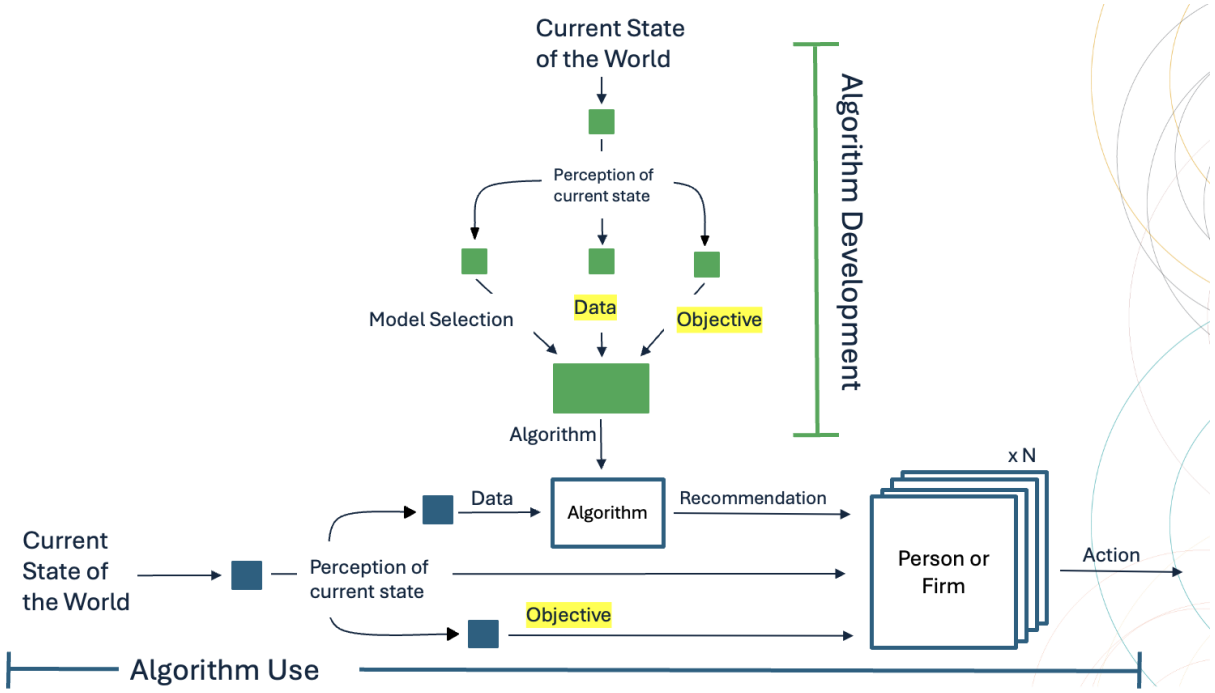

Figure 6.

Kleinberg et al. show that commonly selected *objective functions* result in algorithms that replicate human errors. We can begin to address this problem by designing more effective *objective functions* - ones that take into account findings from behavioral science and psychology, particularly models of human behavior that account for errors in human judgement (Kleinberg et al., 2023). For example, an algorithm that attempts to optimize user satisfaction but uses time engaged as a proxy for satisfaction due to *data* limitations fails to account for the fact that people imperfectly optimize for satisfaction; consider the fact an individual may spend a lot of time on a post that angers them and then regret it later (ibid.). Ideally, an algorithm optimizing user satisfaction would account for the ways in which time spent on a post may not fully correlate with user satisfaction; but at present many algorithms fail to do so (ibid.). The authors introduce the concept of "inversion problems" to clarify the issue. Inversion problems are situations in which we do not have direct data on the object we care about, so must deduce the true object, often a mental state like satisfaction level or knowledge, from our available data - we attempt to "invert" the true mental state we care about from behavioral data. Importantly, the data we have is an imperfect proxy for the object we care about (ibid.). For example, we care about a Doctor's expert judgement, but the data we have - a doctor's diagnosis - includes their expert judgement modified by a range of factors well documented in behavioral science literature, like their level of fatigue (ibid.). Kleinberg et al. advocate for collaboration between behavioral science and machine learning research communities to produce algorithms that address, rather than replicate, flaws in human judgement.

Examining model selection

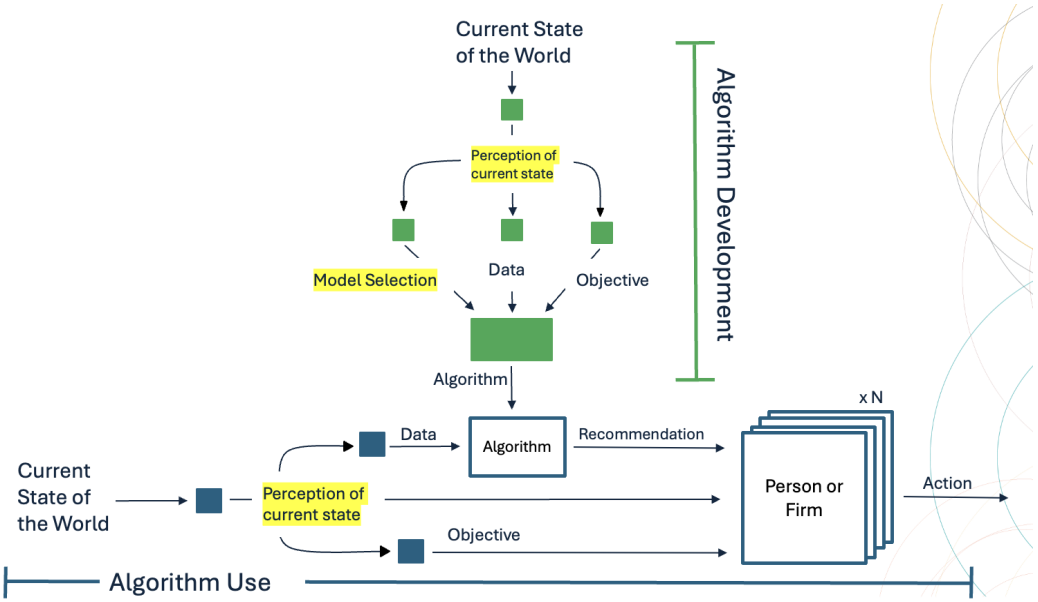

Figure 7.

In an important paper, Li et al. demonstrate the importance of *model selection* to the overall effectiveness of an algorithm. In cases where *perception of the current state* may be limited, significant welfare gains arise from selection of a model that not only optimizes an objective based on known information, but expands the information that is known through targeted exploration. Supervised learning algorithms commonly used for resume screening do not incorporate such exploration. The authors compare the performance of one such supervised learning algorithm to an algorithm that incorporates exploration - a stochastic reinforcement learning algorithm (Li et al., 2020). In a resume screening trial for a US Bank, the authors show that the reinforcement learning algorithm selects an application pool that is comparably successful to the commonly used algorithm and more successful than human decision makers. In the study, "successful" is defined by the Bank's own evaluation criteria (Li et al., 2020). Importantly, the candidate pool selected by the reinforcement learning algorithm is significantly more diverse across gender, race, income, and academic background (eg. major, university attended) than the candidate pool selected by human decision makers or the supervised learning algorithm (ibid.). An additional benefit of the candidate pool selected by the reinforcement learning algorithm is that on average, the candidates offered a position from this pool were more likely to accept the offer than those in candidate pools created by the other selection processes (ibid.). For many companies, this is a significant benefit as recruiting a candidate who ultimately rejects an offer is an inefficient use of resources (ibid.). One hypothesis that explains the difference in candidate pools' acceptance rate relates to our earlier discussion of algorithmic monoculture. Consider the possibility that most of the Bank's competitors use supervised learning algorithms - perhaps the same one, perhaps similar ones - these Banks will compete

over the same candidates (ibid.). For any given Bank, this lowers the chance of a candidate they select accepting their offer (ibid.). Adopting a reinforcement learning algorithm in this context provides a distinct advantage - as one of the only Banks with a different (but comparably accurate) candidate ranking, a Bank faces significantly less competition when recruiting candidates (ibid.). The higher acceptance rate Li et al. observe in the candidate pool selected by their reinforcement learning algorithm may reflect this dynamic, though more research is necessary to confirm this hypothesis.

Evaluating models along multiple dimensions to identify pareto improvements

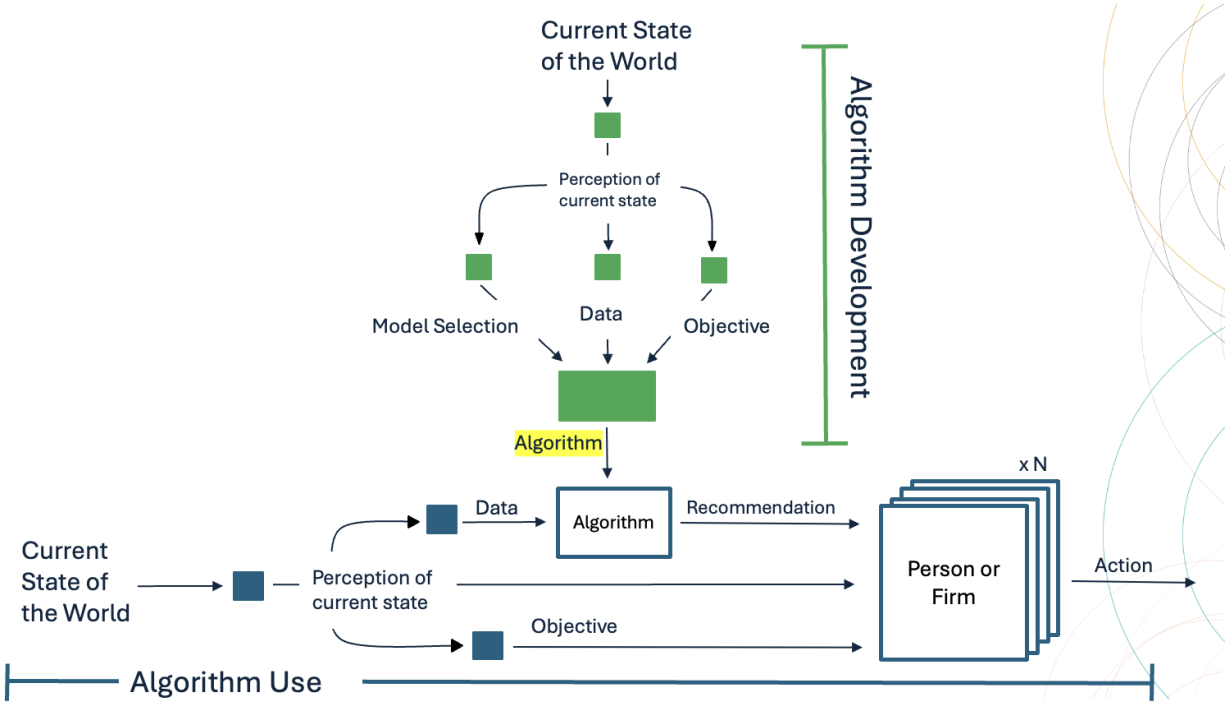

Figure 8.

Algorithms with the same overall performance accuracy may differ significantly when evaluated on other performance metrics, such as fairness (Coston et al., 2021). For example, consider two algorithms with the same average accuracy across a population. One algorithm might have equal performance across most or all demographic groups, whereas the other might have terrible accuracy for, say, women aged 20-60 and slightly higher accuracy across all remaining demographic groups. Coston et al. formalize and instrumentalize this observation - they begin by highlighting the significance of the problem by considering algorithm development in the context of the "Rashomon Effect". The Rashomon Effect indicates that for a given algorithm, often there may be a number of alternate algorithms with "similar overall performance but very different individual predictions" (ibid.). Failure to consider this option set can result in substantial opportunity cost. The authors develop an algorithm that compares fairness across the Rashomon set. With this tool, it is straightforward to replace a less fair algorithm with a more fair one with

no loss, resulting in strict improvement on the benchmark algorithm (ibid.). This tool fits very well into the US legal framework - in discrimination law, disparate impact holds that an algorithm that has no business justification and disproportionately harms members of a protected group (eg. a group defined by gender, age, or race) is not legal (ibid.). Coston et al.'s algorithm identifies situations where this is the case, because there is no business justification for an algorithm that is less fair than one with identical performance (ibid.).

Surprising results from out of sample model evaluations

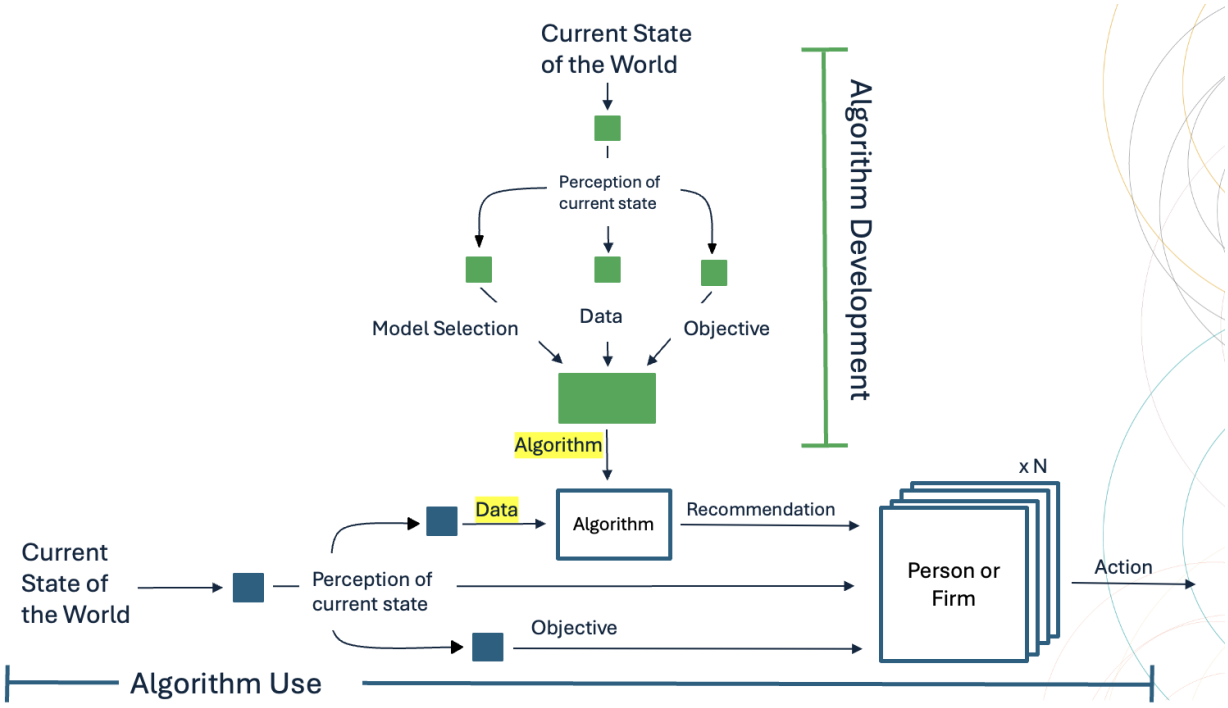

Figure 9.

To ensure algorithmic fairness, greater attention must be paid to algorithms' out of sample performance (Yang et al., 2024). Out of sample in the context of healthcare refers to an algorithm's use once it has been deployed, as opposed to in validation or training. Bias in ML diagnosis tools is well-documented and harmful (ibid.). Yang et al. find that AI diagnostic tools for radiology, dermatology, and ophthalmology are inferring demographic information and using this as a "shortcut" for diagnosis - where the authors define "shortcuts" as a decision based upon "correlations that are present in the data but have no real clinical basis, for instance deep models using the hospital as a shortcut for disease prediction". And, the authors show that use of such shortcuts is correlated with harmful disparities in diagnostic accuracy across demographic groups defined by race, sex, and age. This work takes place in the context of radiology, dermatology, and ophthalmology; the authors train and evaluate models using labeled chest x-rays, dermatology and ophthalmology images. Yang et al. define bias in terms of disparities in false positive and false negative rates across demographic groups. Importantly, and surprisingly, Yang

et al. show that debiasing methods to reduce disparities across demographic groups do not always generalize as expected. They find that a "debiased" *algorithm* may actually still be biased in the context it's used - often due to a *data* distribution shift. A data distribution shift occurs when aggregate characteristics of a dataset - for example, a population mean - change. The authors test algorithms out of sample on data that exhibit various distribution shifts; for a given demographic group, distribution shifts between in sample data and out of sample data may include, among other shifts, different prevalence of diagnoses or different prevalence of covariates. The authors find that often, the model with least bias in sample is not the model with least bias out of sample, however they identify an in-sample debiasing model selection technique that selects models with consistent out of sample bias results. This debiasing approach involves selection of models with minimal encoding of demographic attributes, as defined by 'Minimum Attribute Prediction Accuracy', as opposed to selection of models with maximal in sample fairness (ibid.). Additionally, the authors evaluate the relative performance of three types of debiasing techniques; "methods which remove demographic information from [model] embeddings", methods which "reweight samples based on their group to combat underrepresentation" and methods which "more generically attempt to improve model generalization— that is, exponential moving average" (ibid.). Debiasing techniques involving the removal of demographic information from model embeddings were the most effective when tested out of sample (ibid.). The success of the author's model selection approaches that minimized embedded demographic information provides further evidence that existing disparities in accuracy across demographic groups are influenced by flawed "shortcutting".

Mitigating bias driven by the current state of the world

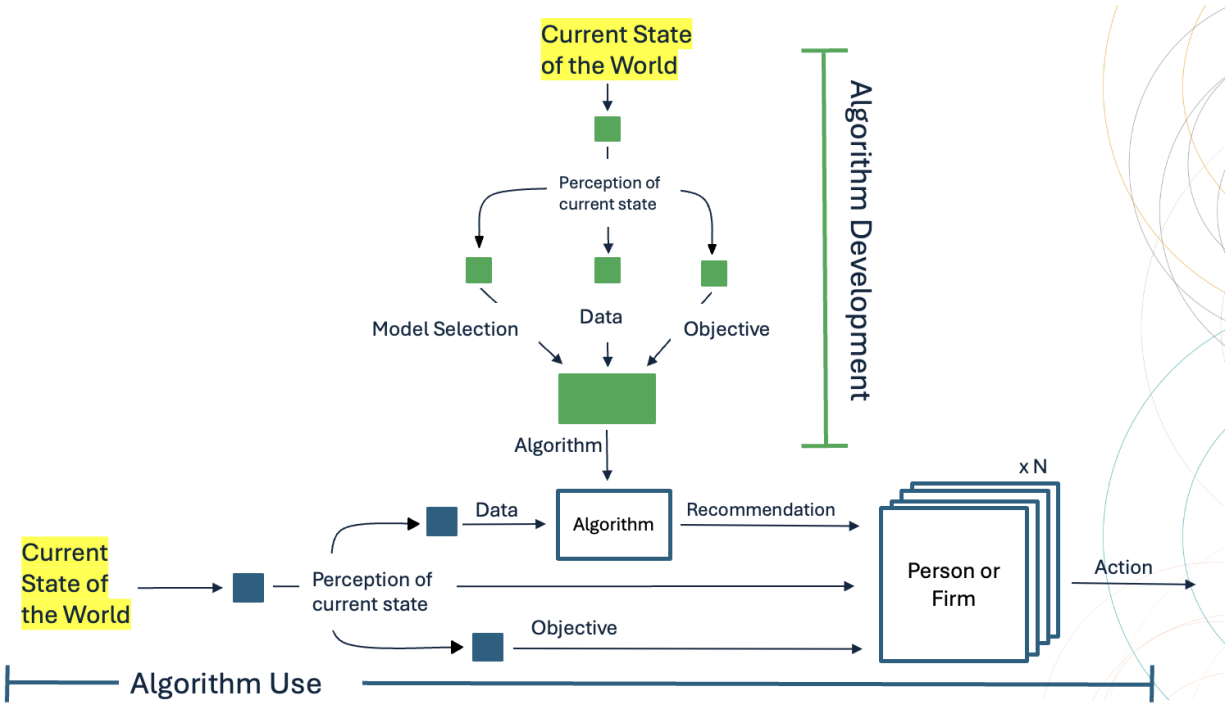

Figure 10.

Without careful consideration of the current state of the world, harmful bias will be overlooked and perpetuated. For example, without intervention, data produced from our *current state* will lead to algorithms that reproduce systemic racism. Consider another example; Lambrecht and Tucker show how an advertising campaign for STEM jobs that is gender-neutral on its face will post fewer ads to women than men. Online ad placement is determined by an auction mechanism - firms submit bids to show their advertisement to a particular user, and the firm with the highest bid receives the ad placement. In the STEM advertising example, even if the amount the STEM company bid on ads to men and women was equal, more ads would be distributed to men than women (Lambrecht and Tucker, 2018). This problem is a function of the *current state of the world*; particularly, spillover effects from firms who place higher bids on ads for women than men, making bidding on ads for women more competitive than bidding on ads for men (ibid.). The consequence of this bidding pattern is that a supposedly neutral advertising strategy, such as optimizing for cost-effectiveness, will result in significant disparities in distribution across gender (Lambrecht and Tucker, 2018). The *current state of the world* also restricts possible remedies to this problem, including implementation of a campaign that shows the same number of ads to women as men (ibid.). Such a campaign would require consistently higher bids on ads shown to women than to men, but certain advertisers, including those advertising jobs, are forbidden from introducing separate bidding strategies based on demographics (ibid.). Though this policy is important and intended to prevent discrimination, it removes an option that would reduce advertising disparities between men and women (ibid.). Greater consideration for the *current state of the world* could improve policy design - for example by introducing opportunities for firms currently prohibited from targeted advertising to engage in it when this

would increase social welfare (ibid.). This might look like giving firms the opportunity to select an advertising strategy that will show ads equally across particular groups, even when this produces different bidding patterns across the same groups (Lambrecht and Tucker, 2018). What we choose to measure plays a significant role in the challenges described in this example - in advertising, current policy is tied to a measurement of bid price, although policy could have instead focused on a measure of advertisements shown (ibid.). A policy that equalizes price bid for ads to men and women results in a disparity in ads shown to men and women. Shifting the focus of our policy from measure of bid price to a measure of ads shown is another solution to the problem described above (ibid.).

The importance of data selection

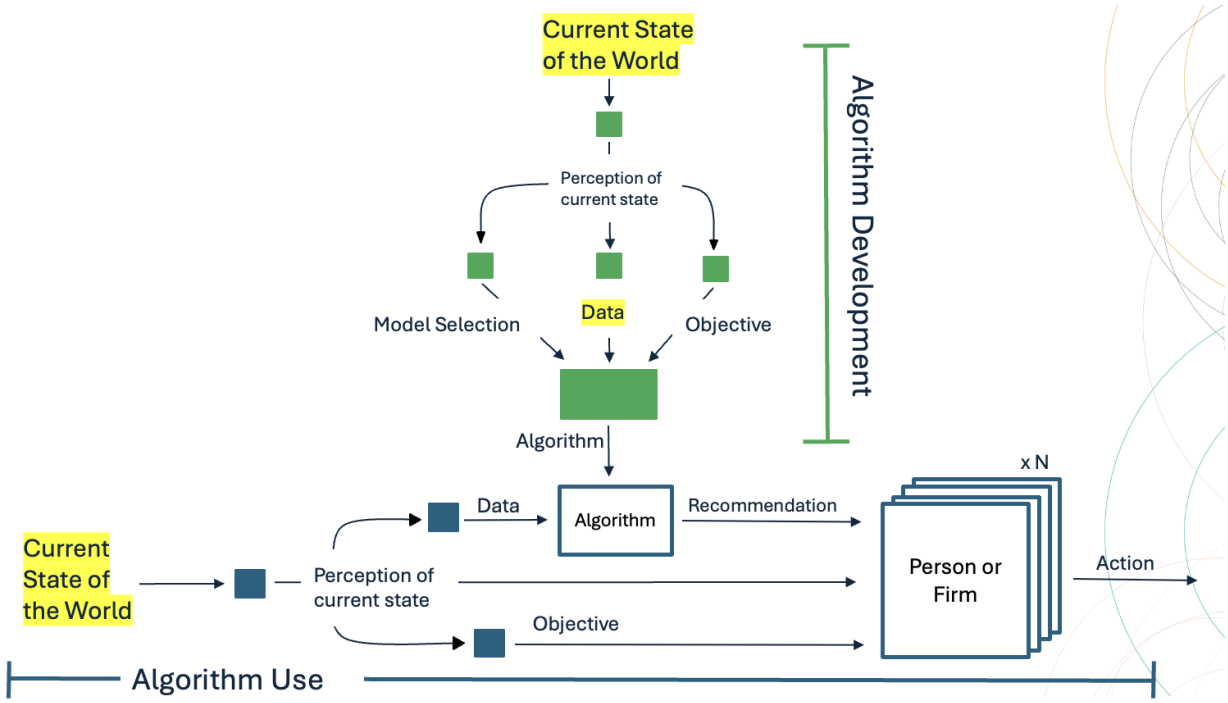

Figure 11.

In “Discrimination in the Age of Algorithms”, Kleinberg et al. develop a framework for identifying and addressing discrimination by algorithms. One of the examples they discuss represents the importance of considering training *data* selection in the context of the current state of the world; this can determine an algorithm’s fairness and accuracy. The authors note that including race in training data may result in fairer and more accurate algorithmic outcomes than those from an algorithm trained on data that excludes such information. This is because given the *current state of the world*, in which systemic racism is pervasive, consideration of race can correct prediction bias caused by the disparate impact of racism. For example, due to historical and current racism, on average a wealth gap exists between Black and white people in the United States. Indicators of future success might differ among people of different income status - for

example, lower-income students may not be as likely to engage in as many extracurriculars, because they need to work after school. In this case, while engagement in extracurriculars may be a good predictor of success in College among high income students, it may be a very poor predictor of success in college among low income students (ibid.). Inclusion of demographic data enables optimal predictors of success to be used in each case (ibid.). Of course, care and consideration of the specific problem at hand should be taken to ensure equity when considering inclusion or exclusion of training data. Rather than make a general normative judgement on the subject, our goal here is to highlight Kleinberg et al.'s point that selection of training data can substantially influence the fairness and accuracy of a model. For a deeper and more comprehensive understanding of this topic, we recommend reading "Discrimination in the Age of Algorithms".

Individual overrides of algorithmic recommendations

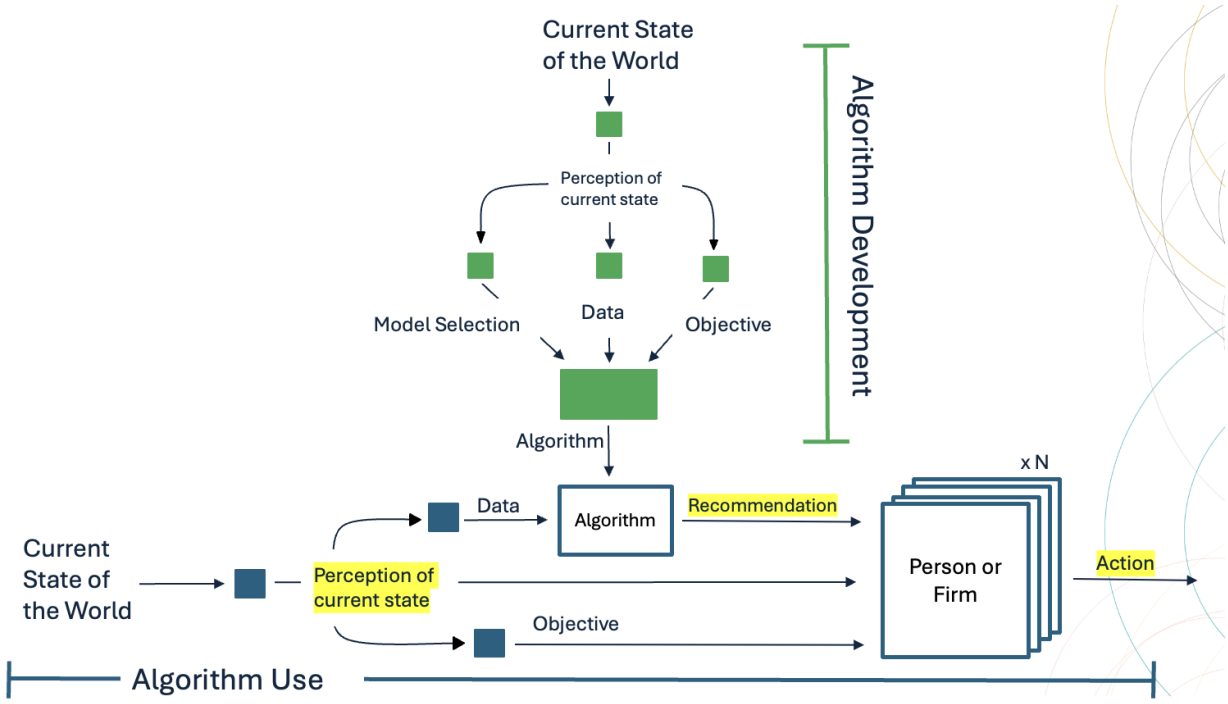

Figure 12.

Pretrial release or jail decisions are extraordinarily consequential; they can be the difference between spending the next five months awaiting trial in jail - and losing one's job - versus at home (Kleinberg et al. 2018). Algorithmic bail recommendations are common, but so are Judge overrides of algorithm's recommendations (Dobbie and Yang, 2023). In an empirical assessment of the circumstances under which Judges override algorithmic bail recommendations, Yang and Dobbie find that for 90% of Judges who override a bail algorithm, their choice results in poorer decisions than the algorithm would have made, while for 10% of Judges who override an

algorithmic recommendation, their choice results in better decisions than the algorithm's recommendation. A significant contribution here is quasi-experimental tools the authors develop to "measure the impact of human discretion over an algorithm on the accuracy of decisions, even when the outcome of interest is only selectively observed, in the context of bail decisions." Judges have a substantial amount of contextual information available to them that is not included in the training or input data to a bail recommendation model (such an algorithm has a selective set of risk factor information, for example the defendant's number of past arrests) (ibid.). Examples of information available to a Judge, but not to an algorithm, include whether a pretrial services officer recommends override of the algorithmic bail recommendation, whether the defendant is homeless, and whether the detailed charge involves violence against adults or children, and whether the current case takes place just after a completely unrelated case where an individual has been arrested for homicide or violent first-degree felony while on pretrial release (ibid.). Yang and Dobbie provide evidence that disparities in how judges use private information, rather than public information (available to an algorithm and a judge) drive differences in effectiveness of algorithmic overrides. They train an algorithm to predict high skill Judge's decisions using only publicly available information. High skill Judges substantially outperform this algorithm, indicating that they are most likely using private information to make their choices (ibid.). Additionally, low and high school Judges appear to use publicly available information in very similar ways (ibid.). The author's work informs designs for human / algorithm interaction and highlights the importance of examining outcomes of algorithmic decision making systems, or *actions* in our model, in addition to outputs of algorithms (*recommendations* in our model).

Conclusion

Our model of AI Decision Making is useful to a regulator, researcher, or practitioner who aims to identify and resolve problems involving AI. Consideration of interdependency in AI decision making systems supports creative and successful problem solving and is necessary for effective regulation. We demonstrate how our model can be useful to understanding and addressing AI problems by mapping onto our model important work that identifies and addresses AI problems. The review of this literature is the second focus of our paper. We hope this paper provides a clearer picture of a complex system and empowers individuals and institutions to change AI Decision Making systems to improve social welfare.

Works Cited:

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology*. NBER Working Paper No. 31422. Cambridge, MA: National Bureau of Economic Research, July 2023.

Angelova, V., Dobbie, W.S. and Yang, C., 2023. *Algorithmic Recommendations and Human Discretion*. Working Paper No. 31747, National Bureau of Economic Research. Available at: http://www.nber.org/papers/w31747 [Accessed 23 Apr. 2025].

Kahneman, D., Sibony, O., & Sunstein, C. R. (2022). *Noise: a flaw in human judgment.* First Little, Brown Spark paperback edition. Little, Brown Spark.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. HUMAN DECISIONS AND MACHINE PREDICTIONS. Q J Econ. 2018 Feb 1;133(1):237-293. doi: 10.1093/qje/qjx032. Epub 2017 Aug 26. PMID: 29755141; PMCID: PMC5947971.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R Sunstein, Discrimination in the Age of Algorithms, *Journal of Legal Analysis*, Volume 10, 2018, Pages 113–174, https://doi.org/10.1093/jla/laz001.

Lambrecht, Anja and Tucker, Catherine E., Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018). Available at SSRN: https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260

Yang, Y., Zhang, H., Gichoya, J.W. *et al.* The limits of fair medical imaging AI in real-world generalization. *Nat Med* 30, 2838–2848 (2024). https://doi.org/10.1038/s41591-024-03113-4

A. Coston, A. Rambachan, and A. Chouldechova. Characterizing fairness over the set of good models under selective labels. In Proceedings of the 38th International Conference on Machine Learning, pages 2144–2155, 2021.

Lina M. Khan, *Amazon's Antitrust Paradox*, 126 YALE L. J. 710 (2017).
Available at: https://scholarship.law.columbia.edu/faculty_scholarship/2808

Li, Danielle, Lindsey R. Raymond, and Peter Bergman. *Hiring as Exploration*. NBER Working Paper No. 27736. Cambridge, MA: National Bureau of Economic Research, August 2020.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Raghavan, M. (2023). The Inversion Problem: Why Algorithms Should Infer Mental State and Not Just Predict Behavior. Perspectives on Psychological Science, 19(5), 827-838. https://doi.org/10.1177/17456916231212138 (Original work published 2024)

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342. PMID: 31649194.

Lina M. Khan, *The Separation of Platforms and Commerce*, 119 COLUM. L. REV. 973 (2019). Available at: https://scholarship.law.columbia.edu/faculty_scholarship/2789

Federal Trade Commission. *Complaint: FTC v. Amazon.com, Inc.* No. 2:23-cv-01495-JHC. Document 114. November 2023.

Farronato, Chiara, Andrey Fradkin and Alexander MacKay. 2023. "Self-Preferencing at Amazon: Evidence from Search Rankings." *AEA Papers and Proceedings*, 113 : 239–43.

Kleinberg, Jon & Manish Raghavan, Algorithmic monoculture and social welfare, Proc. Natl. Acad. Sci. U.S.A. 118 (22) e2018340118, https://doi.org/10.1073/pnas.2018340118 (2021).

Stuart Russell, Karine Perset, & Marko Grobelnik. 2023. "Updates to the OECD's definition of an AI system explained." *OECD.AI*. Available at: https://oecd.ai/en/wonk/ai-system-definition-update.