Philosophy & Technology (2024) 37:68
https://doi.org/10.1007/s13347-024-00757-5

RESEARCH ARTICLE

Algorithmic Decision-Making, Agency Costs, and Institution-Based Trust

Keith Dowding[1] · **Brad R. Taylor**[2]

Received: 23 April 2023 / Accepted: 14 May 2024
© The Author(s) 2024

Abstract
Algorithm Decision Making (ADM) systems designed to augment or automate human decision-making have the potential to produce better decisions while also freeing up human time and attention for other pursuits. For this potential to be realised, however, algorithmic decisions must be sufficiently aligned with human goals and interests. We take a Principal-Agent (P-A) approach to the questions of ADM alignment and trust. In a broad sense, ADM is beneficial if and only if human principals can trust algorithmic agents to act faithfully on their behalf. This mirrors the challenge of facilitating P-A relationships among humans, but the peculiar nature of human-machine interaction also raises unique issues. The problem of asymmetric information is omnipresent but takes a different form in the context of ADM. Although the decision-making machinery of an algorithmic agent can in principle be laid bare for all to see, the sheer complexity of ADM systems based on deep learning models prevents straightforward monitoring. We draw on literature from economics and political science to argue that the problem of trust in ADM systems should be addressed at the level of institutions. Although the dyadic relationship between human principals and algorithmic agents is our ultimate concern, cooperation at this level must rest against an institutional environment which allows humans to effectively evaluate and choose among algorithmic alternatives.

Keywords Algorithmic decision-making · Principal-agent Theory · Asymmetric Information · Explainable AI · Institution-based Trust

✉ Brad R. Taylor
brad.taylor@usq.edu.au

[1] School of Politics and International Relations, Australian National University, Canberra, ACT, Australia

[2] School of Business & School of Humanities and Communication, University of Southern Queensland, Springfield Central, QLD, Australia

Published online: 27 May 2024

1 Introduction

Recent increases in data availability and computing power have rapidly expanded the use of artificially intelligent (AI) algorithms to assist or automate human decision-making (Kochenderfer et al., 2022). Although AI use in decision-making can potentially remove human bias from important business and government decisions (Lepri et al., 2018), recent research shows that algorithms can introduce their own bias (Kordzadeh & Ghasemaghaei, 2022). Algorithmic bias often reinforces existing inequalities and power relations (Gerdon et al., 2022). When a police force uses historically biased crime data to make policing decisions, for example, the costs are primarily borne by the targets rather than the users of the algorithm. We label such harms *algorithmic externalities*: harms caused by the use of algorithms borne by individuals who play no part in choosing whether or how the algorithm is deployed.

Work on algorithmic externalities is extremely important, but our concern in this paper is narrower. We focus on what we label *algorithmic internalities*: harms caused by algorithms borne by the algorithmic users who have at least some say on whether or how the algorithm is adopted. This includes individuals choosing a specific algorithm to assist their own decision-making or customers choosing to use a product or platform which uses algorithms to assist or automate their choices. Such harms are internal to the relationship between the user and the algorithm, rather than being borne by a third party. We restrict our focus in this way not because algorithmic externalities are unimportant, but because (1) the internal dynamics between users and algorithmic agents has received far less attention, and (2) such dynamics bring into relief important considerations around opacity and agency which may later be applied to broader analyses of algorithmic ethics.

In this paper, we analyse algorithmic internalities through the lens of Principal-Agent (P-A) theory (Dowding & Taylor, 2020, pp. 73–78). Principal-agent models consider the incentive problems arising through informational asymmetries in task-delegation. In general, when one person (the principal) delegates authority to another (the agent), the former necessarily is ignorant of the precise endeavours of the latter. Asymmetric information is inherent to principal-agent relationships and cannot be avoided without costly oversight procedures. The P-A problem arises as agents can hide their true abilities (they lack the skills they advertise), might shirk (neglect their duties) or pursue interests other than those of the principal. A car mechanic, for example, might not be competent to deal with a particular type of engine, spend more time on social media than working, or perform unnecessary work to line their own pockets. We examine the relationship between human principals and algorithmic agents using this framework. While the human-machine and the human-human principal-agent problems are not identical, literature on the human-human relationship can provide valuable analytic traction. However, we examine the differences between human and algorithmic agents as these lead to some important peculiarities in the human-machine P-A relationship.

Our analysis indicates that human-machine P-A problems are best addressed at an institutional level by considering the mechanisms through which useful information and judgments are produced and disseminated. Such institutions cannot be expected to perfectly align the actions of algorithmic agents with the interests of their human

principals. Rather, we should seek to create conditions which enable individuals to reasonably choose whether or not to employ an algorithmic agent in a particular context and to meaningfully evaluate the alternative options.

2 Algorithmic Decision-Making and Human Capabilities

2.1 Algorithmic Decision-Making

By algorithmic decision-making (ADM), we mean the use of algorithms to assist or to make decisions on behalf of humans. This ranges from the filtering of choices in recommender systems to fully autonomous decision-making systems. Such algorithms generally analyse large amounts of data to discover underlying patterns and make predictions about the best decision in terms of the user's preferences (usually as judged by past their behaviour or by that of similar individuals). ADM is extensively used in financial trading, medical diagnosis, and many other applications (Kochenderfer et al., 2022).

An ADM system makes decisions much faster than humans while drawing on far more data. Algorithmic decisions also potentially avoid the cognitive biases characterising human decision-making. Algorithms also come with their own biases, however, raising questions about the extent to which ADM serves human interests (Gerdon et al., 2022; Johnson, 2021).

Some ADM systems are relatively simple rule-based systems. Machine learning systems, on the other hand, are often complex. Rather than being programmed with specific instructions, machine learning allows algorithms to learn how to perform some task based on examining a large number of examples, identifying patterns in this training data, and altering itself to produce more accurate outputs (Mohri et al., 2012). This approach has enabled recent advances in generative AI, allowing machines to perform a number of tasks such as image and text generation, surprisingly well.

These recent advances demonstrate what Halevy et al. (2009) call "the unreasonable effectiveness of data." Modifying Eugene Wigner's (1995) observation that mathematics is "unreasonably effective" at explaining physical phenomena with simple equations, Halevey et al. argue that data has a similar power. The success of machine learning applications depends primarily not on the cleverness of the design, but on the availability of massive quantities of data: "invariably, simple models and a lot of data trump more elaborate models based on less data" (Halevy et al., 2009, p. 9).[1] OpenAI's ChatGPT, for example, is able to generate human-like text because it accesses an enormous dataset of written text and given enough computing power to analyse the patterns in this data to build a workable model of how words, sentences, and concepts fit together. The specific calculations it uses are simple, but there are a lot of them. When deciding what word to write next, ChatGPT performs some

[1] The difference is that simple equations summarise what might seem to be mysterious relationships, but data-heavy algorithms do not summarise relationships but model them in situ so to speak. In that sense, they do not seem to explain, but rather predict (Dowding & Miller, 2019).

175 billion calculations – one for each connection in its neural network (Wolfram, 2023). Since these algorithms operate at such a massive scale, human minds are not equipped to comprehend precisely what goes into any particular case and even the system designers do not know exactly how it works (Burrell, 2016). We return to this opacity issue in Sect. 3. However, immediately the asymmetry of information between the machine-agent and the human-principal is obvious.

The level of human involvement is important when considering the social impact of ADM. Ivanov (2023) classifies automated decision-making systems depending on whether humans are "in the loop," "on the loop," or "out of the loop." When humans are "in the loop," the algorithm makes a recommendation, but the decision is ultimately made by the human. Where the algorithm is empowered to decide but the human can override, humans are said to be "on the loop." Humans are "out of the loop" when algorithmic decisions cannot be overridden on a case-by-case basis.

Combining human and machine intelligence can often produce better outcomes (Johnson et al., 2022). For example, Hekler et al. (2019) find that although convolutional neural networks outperform dermatologists in classifying images of suspected skin cancers, combining the judgements of humans and machines produces even greater accuracy. In this case, there is good reason to prefer a high level of human involvement in decision-making. Diagnosing skin cancer is clearly high-stakes, so scarce human time and effort is well-spent improving outcomes. In other contexts, however, the case for human involvement in ADM is weaker. Some decisions are sufficiently unimportant that human involvement is not worthwhile even if it would improve outcomes. In other cases, human involvement decreases the quality of decisions if people overturn optimal choices due to cognitive bias or overconfidence. Kleinberg and Verschuere (2021), for example, find machine learning more accurately detects deceptive statements than humans and allowing humans to overrule machine judgements reduced accuracy.

Different decision contexts call for different combinations of human and machine input, and so general judgements about the nature and scope of human-machine collaboration are unlikely to be cogent. What is clear, however, is that given the current state of the art, there are at least some choices that users can reasonably hand off to machines to produce better or sufficiently good decisions with less human effort. As the technology develops, the number of such cases will grow. We make no specific assumption about the currently optimal level of ADM or the pace of change over time, but we do insist that the cases in which significant algorithmic involvement and even the automation of decision-making can improve outcomes are sufficiently plausible at reasonable time horizons to merit careful analysis.

2.2 ADM and Human Capabilities

We use Amartya Sen's (1999) capability approach to consider the potential benefits of algorithmic assistance in decision-making to human users. Rather than using the actual welfare benefits or the simple rights to judge systems, Sen argues we need to examine how systems increase a person's capabilities. So we do not make judgements on what people actually achieve, say through the education system but what that system enables them to achieve. This involves setting an educational system

that maximises each person's capabilities given that people have different abilities. An individual with attention deficit hyperactivity disorder (ADHD), for example, may have diminished capabilities relative to a neurotypical peer when it comes to education despite having access the same level of resources. Medication and assistive technologies can be used to expand capabilities in education (Black & Hattingh, 2020). Fitting the system to their needs can maximise each person's capabilities thus representing real or substantive freedoms.

ADM could enhance human capabilities in a variety of ways. Algorithms provide external choice-making machinery, potentially reducing the costs of decision-making and increasing the quality of decisions. Excessive alternatives in opportunity sets can reduce well-being (Schwartz, 2009). Meaningful choice requires the choice maker to weigh up the value of alternatives. Excessive numbers of alternatives may force people to "pick" rather "choose" (Ullmann-Margalit & Morgenbesser, 1977). Expanding a library of available movies has an ambiguous impact on the substantive degree of choice as the greater number and diversity can outstrip our ability to rationally choose. Algorithms could enable us to reap the benefits of greater choice without falling prey to the paradox of choice. Algorithms could thus be part of an "extended mind" with technology augmenting the internal components of our decision-making machinery (Clark & Chalmers, 1998; Smart, 2017).

Furthermore, algorithmic assistance can increase welfare by enabling better choices. Algorithms can use vast amounts of information to align with our preferences revealed through past choices. They can be used to create subsets within the vast array allowing us to make the final decision and be less open to the sorts of framing effects that humans are in terms of choice architecture (Thaler & Sunstein, 2021). Humans can also use algorithms to privilege their "higher-order" over reflexive "first-order" preferences (Frankfurt, 1971) overcoming habit or weakness of will (Mele, 1992). For example, if we want to watch more classic films, we could alter our streaming algorithm to privilege such options. An aspiring gourmet could choose an algorithm emphasising complex flavours; a weight watcher other dietary requirements.

In this manner ADM enables the adoption of a disposition to choose rather than the final choice itself. We rationally commit to a future pattern of behaviour through the alteration of the process by which we make day-to-day decisions (Brennan & Hamlin, 2008; Hamlin, 2006). ADM can potentially alter our decision-making machinery, saving decision costs, making use of more information, and enabling us to commit to what we think are better choice dispositions. The question becomes how this happy outcome can be enabled without creating greater dangers of losing our ability to choose or be dominated by our enhanced decision-making machinery.

3 Principal-Agent Theory

While AI algorithms may augment human choices, they are also opaque. We want to ensure that algorithms faithfully serve human aims and interests, but informational asymmetry makes this problematic. Such agential asymmetry is analysed by principal-agent (P-A) theory (Dowding & Taylor, 2020, pp. 73–78). There are several

elements to P-A problems, but the central issue pertinent to algorithms is asymmetric information when the interests of the agent and the principal diverge. Grounded in economics, P-A theory typically assumes both principal and agent are rational and self-interested. The diverging interests and information asymmetry leads to agent *opportunism*. The principal cannot monitor and punish agents pursuing their own interests without excessive costs. The challenge faced in P-A models is to structure incentives to minimise agency costs.

A familiar P-A example is that between an auto mechanic and customer. The customer wants their car serviced as well and cheaply as possible. The mechanic wishes to maximise profit. Since the customer cannot perfectly observe the actions of the mechanic, there is a risk of opportunism. The mechanic may misrepresent costs, undertake unneeded repairs, use substandard techniques or materials. Information asymmetry makes it difficult for the customer to do much about this even if they suspect malfeasance.

The customer may attempt to reduce informational asymmetry in various ways, but since the entire point of P-A relationships is for the principal to utilise the expertise or knowledge of the agent this tends to be self-defeating: "by definition the agent has been selected for his specialised knowledge and therefore the principal can never hope completely to check the agent's performance" (Arrow, 1968, p. 538). If a principal can perfectly monitor, they might as well do the task themselves. This remains true even when there is no difference in expertise. A skilled engineer hiring a car mechanic may not lack the relevant knowledge but is still reliant on the mechanic looking under the hood. If the principal requires perfect monitoring, they might as well do the job themself.

We summarise the standard human-human agency relationship in Fig. 1. The principal delegates to the agent and monitors their performance. The agent performs the task but is incentivised to shirk due to asymmetric information.

Bostrom (2014) discusses principal-agent theory in the context of super-intelligent artificial general intelligence (AGI). Should such intelligences come to exist these considerations will be important, but we set them aside here to deal with contemporary issues of human-machine agency problems which fall short of AGI. We assume

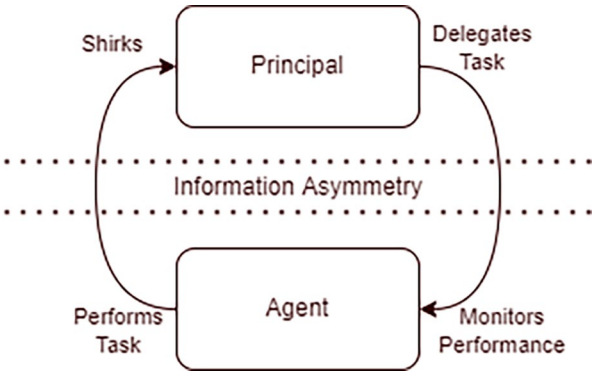

Fig. 1 Principal-agent theory

automated agents have objectives they seek to optimise but are not capable of forming or intentionally concealing their own design-independent preferences. Drawing on the fields of New Materialism and Algorithmic Governance, Kim (2020) argues that algorithms as technological artefacts can be conceptualised as agents. Kim focuses on three P-A relationships: that between civil servants and private companies using algorithms on behalf of government; between corporate managers in private companies and computer scientists; and between computer scientists and the algorithms which are empowered as agents. We build on Kim's analysis similarly conceptualising algorithmic artefacts as agents but focus on a different set of human-machine P-A relationships. We consider human-machine agency relationships between the *user* and the algorithm, rather than the *creator* and the algorithm. We also follow Kim in drawing out asymmetric information in the context of machine learning via Burrell's (2016) analysis of opacity.[2]

Algorithms do not have interests or preferences in the same sense as humans. They are coded to optimise over objectives but lack the underlying psychological features. Hence, when we speak of algorithmic agents' "opportunism" that term is not intended to carry human psychological and moral baggage. The current state of ADM and AI falls short of moral agency, and as such the actions of an algorithm cannot be morally evaluated in the same manner of those of a human. Algorithmic opportunism, however, remains a useful analytic concept insofar as it allows us to capture situations where the decisions made by algorithmic agents diverge from the preferences of their human principals. As explained below, this could result from design choices which intentionally favour commercial over user interests or from unintentional bias in the algorithm.

3.1 Opacity and Imperfect Information

Burrell (2016) distinguishes between three forms of opacity: (1) as intentional secrecy, (2) as technical illiteracy, and (3) as arising from the characteristics and scale of algorithms. We explain each in turn and comparing them to information asymmetries in human-human P-A relationships.

3.1.1 Opacity as Intentional Secrecy

Intentional secrecy, which we call type 1 opacity, occurs when organisations hide algorithmic details. Such secrecy may serve corporate self-interest preventing emulation by competitors or consumer evaluation. It may also serve the more noble goal of reducing manipulation or algorithm gaming. Fully open search engines or spam filtering algorithms, for example, are more manipulable than secret ones (Burrell, 2016, pp. 3–4). This is similar to the information asymmetry dominating human-human P-A relationships. Principals do not have direct access to the thoughts and preferences of agents and can only imperfectly observe their actions. This raises the

[2] Another application of P-A theory to ADM is Borch's (2022) qualitative empirical analysis of human-machine agency relationships in the context of automated trading in financial markets.

prospect of opportunism as the incentives of principals and agents are not perfectly aligned.

The problem of secrecy might be avoided in human-machine relationships via mechanisms inapplicable to human-human ones. We never know with certainty what another human is thinking and might not fully trust them even when they are, in fact, being entirely honest. In contrast, an algorithm can in principle be laid bare by uncovering the code and data. Even if secrecy from the public is necessary or desirable to prevent algorithm gaming, it may be made completely open to trusted third party auditors or regulators in a manner not possible with humans.

It should be noted that this approach of laying the mechanism bare only promotes transparency to the extent that the operation of the system can be understood by humans. As we argue below, this approach is not feasible for sufficiently complex machine learning models beyond the ken of even the most knowledgeable humans. In cases where the mechanistic operation of an algorithm *is* sufficiently well understood, however, transparency (whether direct of mediated by trusted third parties with privileged access) could solve the issue of asymmetric information and machines could credibly to commit to pursuing stated goals. Unfortunately, this only applies to simple algorithmic systems where decision-making operations are encoded in a human-comprehensible form.

3.1.2 Opacity as Technical Illiteracy

Opacity as technical illiteracy (type 2 opacity) can occur even if the code and data of an algorithm are fully open. Just as most of us must trust our mechanic because we do not understand how engines work, most of us do not understand how machine learning algorithms work and must trust the experts in this domain. More general technical training might reduce such opacity, but as AI algorithms grow ever more complex it is unlikely that differences in understanding due to specialised knowledge will come close to being eliminated (Burrell, 2016, p. 4).

3.1.3 Opacity as Complexity and Weirdness

Burrell's (2016, pp. 4–5) concern is primarily with type 3 opacity, which derives from the scale at which algorithms operate. The issue here is that the system is not natively understandable by humans, even those with expertise and deep knowledge of its design:

> *"When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension. Machine optimizations based on training data do not naturally accord with human semantic explanations." (*Burrell, 2016, p. *10)*.

Consider again the 175 billion calculations of ChatGPT when deciding which word comes next when writing. The OpenAI engineers know the system design and can look at the data set it was trained on. In principle they could follow teach of the 175 billion calculations so strictly speaking there are no necessary secrets between

the algorithmic agent and the human principal. The problem is that humans have neither the time nor working memory to follow through these steps. Indeed, following the decision-making process would involve replicating the work in parallel, making the agent's work superfluous. This is precisely the issue seen in human-human P-A problems.

The further problem for the machine case, however, is that the machine might utilise patterns in the data that humans cannot see. The machine categorises in a manner alien to us. The problem is one of interpretation. We cannot interpret what the machine is doing, even if we like the results. In that sense, we cannot be sure that the machine is tracking our desires in terms of the reasons we have for them.

At the most general level, then, the challenge of controlling algorithmic agents appears to be the same as controlling human agents: we must find ways of ensuring the agent is acting in our interest even though we cannot know in detail how and why it is making the choices it does. Type 3 opacity is unique to human-machine relationships, however, and this fundamentally changes the nature of the information problems which arise.

3.2 Human-Machine P-A Relationships

To focus on the problems arising specifically from human-machine relationships, we first concentrate upon type 3 opacity where users can view the algorithm's code and data and have a high degree of technical knowledge about how such algorithms work. The complexity and weirdness of algorithmic decision-making, however, prevents the user from fully understanding how and why decisions are made. We reintroduce the other two forms of opacity later.

Unlike human-human P-A relationships where asymmetric information causes problems, it is the absence rather than inequality of information which can cause divergence from the agent's action to the principal's interests. The issue becomes important if the information that the machine is using does not track our conscious reasons for past choice, but rather reasons due to framing or other non-reasoned causes. A machine might find a non-conscious pattern in our previous choices. For example, when we thought we were choosing exciting action and adventure films, a better predictor is films with pictures of flames, smoke, or damaged buildings. Items that correlate, but not perfectly with action movies. Or the algorithm might pick out the fact we tend to choose the middle items rather than early or late items. While these rather simple examples might be easily overcome, deep learning models will often discover patterns we cannot understand so cannot design against. Even where the algorithm can more accurately represent the preferences of the user it might not be choosing by the agents' own reasons. Over a specific set of items divergence in reasons might not matter for actual choice, but as the set expands over a new range of items, the reason-divergence could lead to choice-divergence. The agent will not be choosing what the principal would have chosen.

These problems are represented in Fig. 2. Like human-human relationships, the user as principal delegates tasks to the algorithm as agent and must monitor the degree to which performance aligns with their preferences. The "shirking" in this case is driven not by selfishness but by the unavoidable discrepancy between the

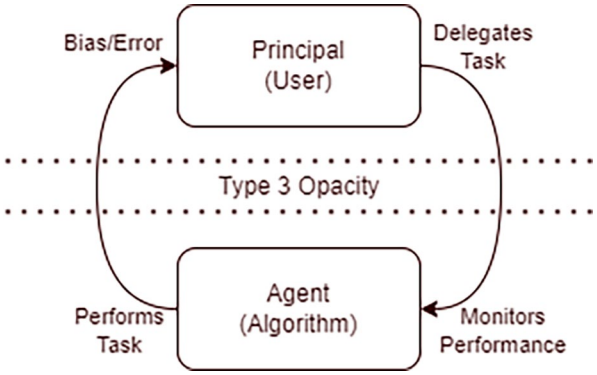

Fig. 2 Human-machine P-A relationships

user's true preferences and the agent's representation thereof. The monitoring challenge is to ensure that the machine learning model is well calibrated to the objectives of the task and to take appropriate corrective action if not.

It is worth reiterating that complete elimination of the information asymmetry is not viable. We employ algorithmic decision-making agents precisely so that they can draw on information we do not have the ability or inclination to use effectively. Just as it would be self-defeating to learn everything our mechanic knows in order to monitor perfectly, it would be self-defeating (and for any reasonably complex model simply impossible) to micromanage algorithmic agents to such a degree that they do not possess any information we lack. The difference is that we can interrogate human agents and, while we might not be able to understand all the details, we can at least attempt to follow the reasoning. To the extent that this is not possible with a machine agent, this is another level of information asymmetry.

Although perfect information is not attainable, some degree of transparency in algorithmic decision-making is widely seen as an important pre-condition for trust (Papantonis & Belle, 2023). There are various approaches to designing transparency in ADM systems (Belle & Papantonis, 2021; Hoffman et al., 2018; Miller, 2019). One approach, known as explainable modelling, is to favour simpler systems which can be interpreted by humans with the relevant technical knowledge (Rudin, 2019). Another, known as mechanistic interpretability, involves reverse engineering complex models to determine the functional decision-processes (Kästner & Crook, 2023). A third approach of post-hoc explanation seeks to provide some sort of human-comprehensible explanation of why a decision was made without providing a comprehensive mechanistic account of the causal chain from input to output (Kenny et al., 2021; Miller, 2019). This approach emphasises the social and psychological dimensions of explanation. Successful explanation involves a transfer of knowledge from one party to another, which depends on contextual factors such as the background knowledge of the audience and their reason for wanting an explanation (Miller, 2019; Sørmo et al., 2005).

Greater understanding of how algorithmic agents reach decisions tends to reduce agency costs but not without other disadvantages. Simpler models are easier to monitor, but for at least some ADM applications more complex models will perform bet-

ter. Designing for both predictive accuracy and interpretability involves optimisation over more than one objective, and conflict between these objectives is to be expected (Jin & Sendhoff, 2008). This suggests that in some cases there will be a trade-off between accuracy and interpretability. This does not imply that complex models are *always* or even *generally* more accurate (Rudin, 2019). Nonetheless, the fact that an accuracy-interpretability trade-off *sometimes* exists means that demanding interpretability requirements would preclude some valuable uses of ADM. Similarly post-hoc explanations falling short of full mechanistic interpretation are costly to produce and necessarily incomplete. Explainable AI is therefore not a general solution to the problem of information asymmetry.

3.3 Algorithmically Mediated Human-Human P-A Relationships

Where human-machine agency relationships are mediated by other humans or organisations, other forms of opacity become relevant. Algorithms will often be "double agents" (Andeweg, 2000) simultaneously serving the interests of users as well as technology companies. Moreover, users will enter into a more complex agency relationship requiring them to consider the alignment of both the company and the algorithm to their preferences. Although the nature of the agency relationship may depend on the particular details of a case, Fig. 3 depicts a straightforward example in which a user tasks a company with deploying an algorithm on their behalf and thereafter interacts directly with the algorithm giving a hierarchical set of agency relationships. The user delegates to the company the task of creating and maintaining an algorithm

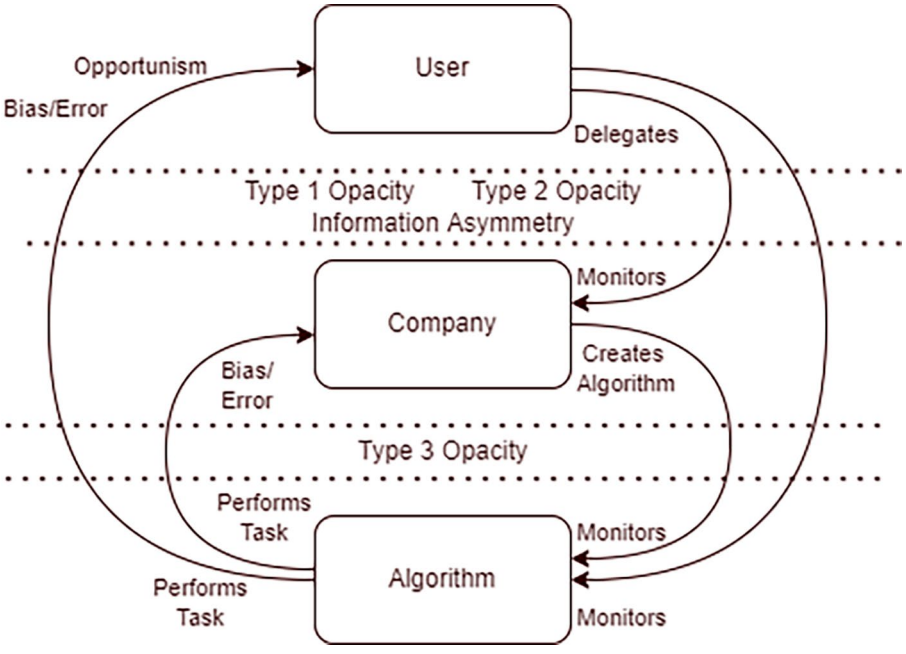

Fig. 3 Algorithmically mediated human-human P-A relationships

(in exchange for payment in money, data, or attention), which in turns creates an algorithm to act as an agent to the company as well as the user. The user monitors the performance of both the company and the algorithm.

This monitoring is compromised by all three forms of opacity. The relationship between user and company is clouded by the first two forms of opacity: secrecy and technological illiteracy. As discussed above, companies normally keep the precise nature of their algorithms secret for various reasons. Most users lack technological knowledge of how machine learning algorithms operate. These two factors lead to a sharp asymmetry of information between users and companies, and as in other human-human agency relationships this can result in opportunism by the commercial company. For example, an e-commerce website might direct customers towards products with higher markups rather than closely matching predicted preferences. This can happen straightforwardly through secrecy, but even without secrecy companies could make strategic use of type 2 and type 3 opacity to create algorithms which prioritise company interests. Again, informational asymmetry is important. Whereas an algorithm is unable to respond to the ignorance of the user by engaging in opportunistic behaviour the companies deploying such algorithms can do exactly that.

Type 3 opacity is present in the relationship between the user and the algorithm, as discussed above, and also the relationship between the company and the algorithm. Although the creators and managers of an algorithm may have direct access to the code and data of an algorithm and be experts in the relevant fields, they nevertheless lack complete knowledge of how and why the algorithm is making its decisions. This leads to a certain amount of slack in the relationship between companies and their algorithms. A perfectly designed algorithm would optimally pursue the objectives of the company. Type 3 opacity, however, will introduce bias and error which will in some cases work against the preferences of users and in others against the preferences of the company.

From the view of the user, the algorithm performs suboptimally due to a combination of the company's exploitation of asymmetric information (type 1 and 2 opacity) as well as bias or error resulting from type 3 opacity. The company as well as the user have incentives to mitigate the effects of such bias and error, but there are conflicting interests regarding opportunism arising from secrecy and technological illiteracy. When an algorithm produces some undesirable outcome from the user's perspective, it is not immediately clear what form of opacity and which agent is to blame.

The monitoring problem described above is serious enough that it might *prima facie* be thought to undermine the possibility of beneficial ADM. If users cannot easily tell when algorithms are serving their interests or where to assign blame when they do not, we may be better off without them. One can make judgements based on outcomes. One can appreciate, at least for simple uses, that life is better or easier or one seems to enjoy the outcomes more using the machine than life was before, before using it. What one cannot judge is whether it could be better still, how far one is being exploited, and whether one is being changed by the algorithm. That might also be so, however with human agents carrying out task for the principal. Although concerns remain, we argue that a shift in focus to the institutional bases of trust represents a promising path forward. If the relevant question is whether we ought to trust the

decisions made by machines, it seems to be *evaluability* rather than *transparency* or *explainability per se* which is most important.

Transparency and explainability are obvious mechanisms for increasing trust and trustworthiness, but they are not the only ones. As in human-human agency relationships, the degree of trust depends on the social, political, and economic environment in which the human-machine agency relationship is embedded. The agency relationship between a customer and a mechanic also presents a significant challenge which would be very difficult to solve through a consideration of only this dyadic relationship, but we often make judgements through public reputation as well as trial and error. We could learn about company algorithms from the press, social media, and acquaintances. Individuals acting severally and collectively can solve many difficult problems by creating appropriate institutions (Ostrom, 1990). We argue that in the context of algorithmic agency relationship, as in others (Rodrik et al., 2004), institutions rule.

4 Institutions Facilitating Choice among Algorithmic Agents

In Sect. 2 we argued that ADM systems have the potential user-benefit of improving the quality of choices notably by aligning with higher-order preferences, freeing scarce human attention for other pursuits. This enhances human capabilities by enabling people to reach otherwise unattainable outcomes. In Sect. 3 we argued that agency costs deriving from asymmetric information could frustrate such potential gains. As in more familiar P-A relationships, excessive agency costs resulting from asymmetric information can discourage the formation of potentially beneficial relationships.

There will be no generally optimal level of algorithmic monitoring or human involvement in ADM systems. Depending on the relative decision-making performance and costs as well as the stakes of the decision, humans should variously be in-the-loop, on-the-loop, and out-of-the-loop (Ivanov, 2023). To fully realise the potential gains of ADM, human principals should be enabled to delegate as much authority to algorithmic agents as appropriate. Moreover, they must be enabled to choose algorithms most closely aligned with their higher-order preferences. This requires what Kim (2020) calls "algorithmic pluralism" – the ability to "make meaningful choices among a multitude of algorithmic decision systems provided by various providers" (Kim, 2020, p. 7). While noting that algorithmic pluralism is important, Kim does not consider the institutional features facilitating it. We propose that algorithmic pluralism requires that human principals are effectively enabled to choose among multiple algorithmic agents based on the extent to which they can be expected to serve the principal's higher-order preferences.

In a basic and limited sense, pluralism can be provided by the familiar institutions of competitive markets and anti-monopoly regulation. If multiple algorithmic systems taking different approaches to ADM exist alongside one another, individuals have greater algorithmic choice. From an economic perspective, we can think about algorithmic pluralism in terms of the degree of competition in the market for algorithmic agents. The standard tools of antitrust policy will be useful (Viscusi et al., 2018),

although the specifics of digital markets (Spulber & Yoo, 2013) and algorithms (Portuese, 2022) raise their own issues for competition economics and policy. The practical challenge of ensuring competition is likely to be significant, but familiar analytic tools and policy instruments are available.

The issue of type 3 opacity further complicates the issue. If principals cannot understand how agents are making choices, how can they determine which are more likely to serve their higher-order preferences? We draw on work from economics and political science to consider the extent to which type 3 opacity problems can be mitigated through appropriate institutions. We do not pretend to offer a complete institutional solution to the issue of human-machine agency relationships. Institutions result from a series of past innovations made by people individually and collectively trying to solve practical problems (Ostrom, 1990). Effective institutions thus emerge in an evolutionary manner and so cannot be specified in the concluding section of an academic paper. Moreover, real-world policy and institutional arrangements have multiple impacts involving complex interactions (Cairney et al., 2019; Durlauf, 2012). We here consider institutional features underpinning meaningful choice among algorithms, but in evaluating real-world institutional alternatives we must also consider other factors. Institutional choice will, as always, be among imperfect alternatives (Demsetz, 1969). Our analysis here aims to provide insights to guide real-world choices, not to offer a comprehensive real-world solution.

New Institutional Economics is grounded in the proposition that appropriately structured institutional arrangements enable economic and social coordination by economising on transaction costs (Williamson, 1985). Consider Coase's (1937) analysis of firms and markets. According to the abstract "blackboard" version of the market economy proposed by neoclassical economists, all economic activity is governed by the price system in a decentralised way. Coase notes that in reality markets are constituted primarily by firms rather than individuals, and within firms central planning, not decentralised exchange, directs production. The explanation is transaction costs. People organise themselves into firms rather than engaging in sets of spot market transactions because the latter means continually finding and negotiating with new trading partners. Crucially for our purposes, the institutional alternatives of markets and hierarchies become endogenous decision variables to the analysis, not immutable background conditions. If both parties to a potential trade see benefit in forming an ongoing contractual relationship rather than keeping their options open, this they can choose (Williamson, 1985, p. 4). Similarly, in human-machine agency relationships. Rather than focusing on the transparency of specific algorithmic decisions within a fixed institutional context, we should consider the institutional environment which embed such relationships. If this environment is structured in a way enabling users to meaningfully choose among algorithmic alternatives, the benefits of ADM can be realised while mitigating the risks.

We distinguish between "dyadic" and "institution-based" trust (Pavlou & Gefen, 2004; Shapiro, 1987). In dyadic relationships, trust can be produced by aligning incentives or through repeated interaction allowing for close and enduring cooperation between two parties. The extent of cooperation and trust we see in large scale industrial societies goes beyond this, however. Handing money to a stranger expecting goods in return requires trust, but when buying a coffee our trust is not based on

the behaviour or characteristics of the barista alone. Rather, we exist in an institutional environment which allows such exchanges to be made without the need for extensive monitoring and enforcement. Fair trading laws, food safety standards, and the degree of competition in the hospitality industry all play into our trust.

As we buy a coffee these institutional factors fall unnoticed in the background, nonetheless they structure and enable consumer choice while incentivising businesses to act in ways consistent with customer interests. These institutions allow us to purchase without being overly concerned with what occurs on the supply side. A cup of coffee is a low-stakes example but in high-stakes examples, such as flying, the institutional trust is more important still, governing all aspect plane manufacture, maintenance and operation of planes and airports.

There is little difference between type 2 and type 3 opacity in dyadic terms. If a principal cannot understand how an aeroplane or ADM system works, the fact that someone else can is relevant only to the extent that such expert knowledge can be drawn upon. This requires institutions. Generally, information problems in markets are mitigated through the use of information intermediaries – third parties facilitating transactions by producing and communicating information to buyers and sellers (Rose, 1999). To the extent information enables better choices, it is valued and therefore commands a market price (Bergemann & Bonatti, 2019; Stigler, 1961). Travel agents, financial advisers, comparison shopping websites, product review publications, and product certification organisations all provide useful consumer information. Reducing asymmetric information increases willingness to buy so principals become more willing to employ agents (Akerlof, 1970; Williamson, 1985).

Since information enables better decisions, agents are willing to pay for it. Third-party information intermediaries such as financial advisors and auditors charge for their services, while product review and comparison-shopping websites monetise consumer attention through advertising. Sellers are also incentivised to offer credible assurances to consumers about quality, safety, and other desirable features. This can be provided by employing independent auditors, certification bodies, or ratings agencies (Klein, 2002). Government regulation also plays a significant role requiring producers to disclose information (Beales et al., 1981) and setting minimum safety and quality standards (Marette et al., 2000).

Information intermediaries and policy interventions do not generally aim to maximise the knowledge of consumers. Often, they provide simple numerical ratings indicating overall quality. Voluntary certification and regulatory standards generally provide pass/fail endorsements of quality or safety. These do not minimise informational asymmetry, rather a third party makes an *evaluation*. If this third party has the requisite information, competence, and impartiality to render a trustworthy judgement, evaluative information can be communicated in a highly efficient way. This allows people to make meaningful choices on the basis on limited information.

Returning to the example of airline safety, this shows why the difference between type 2 (technological illiteracy) and type 3 (complexity and weirdness) opacity remains important in practice. The trust of a passenger is not formed dyadically but institutionally, with the passenger trusting that safety standards for the flight are sufficiently rigorous. The passenger does not understand the complexity involved in every flight, but trusts that institutions are designed to provide a sufficient level of

assurance. In the face of type 1 and type 2 opacity, successful institutional arrangements allow for the production of trust through expert judgment.

With type 3 opacity, however, there are by definition no human experts with such comprehensive understanding. Giving experts unfettered access to the inner workings of an ADM system is of no help if such workings are beyond their comprehension.

Being able to explain the ADM decision might not be essential to the production of trust, however. The challenge for algorithmic choice is not understanding *per se* but rather evaluation. Understanding can contribute to evaluation, but often is not required. If information intermediaries can provide useful evaluations indicating how well an ADM system works, meaningful choice among algorithms becomes possible. The challenge of type 3 opacity, of course, is that even experts cannot comprehend the inner workings of the algorithm so cannot render evaluative judgements. A promising method is to analyse algorithms on the basis of the social scientific audit study method. Audit studies involve applying the method of controlled experimentation to opaque real world processes (Gaddis, 2018). The researcher feeds inputs with randomly varied characteristics into a process observing how input differences result in different outputs. The most commonly referenced audit study is Bertrand and Mullainathan (2004), which audited the American job market for racial discrimination. Employers will not admit such discrimination nor permit researchers to monitor their hiring processes in detail. Simply examining labour market outcomes suffers from confounding factors such as existing patterns of discrimination. In other words, the job market is too complex and opaque to evaluate for fairness on the basis of its internal workings. To test for racial bias avoiding these issues, the authors sent identical fictional resumes to employers varying only the name of the applicant. Those with stereotypically white-sounding names (e.g. Emily and Greg) were 50% more likely to be invited for an interview than those with stereotypically black sounding names (e.g. Lakisha and Jamal).

Numerous algorithm audits have followed identifying bias and discrimination in areas such as hiring algorithms, search results, and facial recognition (Brown et al., 2021; Metaxa et al., 2021). Bias can be detected without looking under the hood, by understanding algorithms "from the outside in" (Metaxa et al., 2021). For example, Buolamwini and Gebru's (2018) *Gender Shades* discovered commercially available facial recognition software was less able to accurately identify female and darker-skinned faces. Algorithm audits tend to focus on algorithmic bias and discrimination, with particular attention paid to algorithmic externalities. Outsider oversight requires an institutional environment providing sufficient independence for auditors and setting professional standards (Raji et al., 2022).

The general approach to understanding algorithms by systematically varying inputs and observing outputs is potentially useful in enabling meaningful choice among algorithms under conditions of type 3 opacity. However, the motivation for such audits and the institutional features required to support them are quite different. Audits for algorithmic externalities have been motivated by activists concerned with social injustice and companies complying with current or prospective regulation (Krafft et al., 2021; Mökander, 2023). On the other hand, the pressure to address algorithmic internalities must come primarily from the human principals themselves. This means that market incentives are more likely to push in the right direction. As

we saw with information intermediaries generally, there is market demand for information to improve decision-making. If consumers of algorithms are willing to pay for information, there is an incentive for the production and communication of such knowledge (Bergemann & Bonatti, 2019; Klein, 2002). Information intermediaries assisting with algorithmic choices could, like financial advisors, provide consultation services for a fee. Alternatively, as in the case of consumer publications, more generalised evaluative information funded through subscriptions or advertising. Such services would likely rely on various sources of information including the audit study approach and everyday explanations of why the system has reached one decision rather than another.

However, the individualised nature of choice among algorithms presents an additional challenge compared to overall assessments of algorithmic bias. If a user wants to assess the degree of alignment between a particular ADM system and their own higher-order preferences, information created for a general audience will be of limited value. Advising for algorithmic choice will also require more careful attention to the *communication* of information than audits aimed at identifying algorithmic externalities. In the latter, the primary audiences (activists, policymakers, target companies) of evaluations will have more time and a higher level of technical knowledge enabling them to absorb the information. In this case, the primary concern of the auditor should be ensuring that the evaluation is accurate. On the other hand, information intended to inform everyday algorithmic choices must be more digestible. Contrast and comparison are also likely to be important when it comes to evaluation and provision. For example, users could be provided with information about what decisions would have been reached by alternative ADM systems.

Market-based information intermediaries could be complemented by non-market institutions. Here, work by political scientists and psychologists on low-information rationality is useful (Hindmoor & Taylor, 2015, Chap. 8). Survey evidence consistently finds most voters lack the degree of political knowledge which might seem necessary for meaningful democratic choice. If voters are not aware what policies are endorsed by each candidate, for example, the extent to which voting decisions represent reasoned judgments may be questioned (Somin, 2016). However, voters appear to rely on a variety of information shortcuts enabling them to make reasoned choices without gathering and processing large quantities of specific political information (Popkin, 1995). Voters base their choices on the opinion of more knowledgeable friends, the endorsement of trusted authorities or their accumulated trust in political parties based on past experience. As with market-based information intermediaries, this enables users to reasonably choose among algorithmic alternatives without imposing excessive informational burdens on them. In other words, algorithmic pluralism promotes "yardstick competition" (Shliefler, 1985), with the users of one ADM system drawing on the experience of those using other systems to comparatively evaluate performance. Such processes become more efficient as we evaluate our and others' experiences over time. This requires a plurality of options to choose from as well as some means of comparative evaluation.

Government regulation has role to play here. Government can produce and disseminate information directly, but more generally by adopting policies supporting choice-enabling market and non-market institutions. Information disclosure require-

ments are an important policy tool (Beales et al., 1981; Vining & Weimer, 1988), and in the context of ADM guarantees of access for third-party auditors (Raji et al., 2022) or the mandated release of basic information and performance metrics (Mitchell et al., 2019) are likely to create more favourable conditions for a robust ecosystem of information intermediaries to emerge.

For the benefits of ADM to be realised, human principals need some way of evaluating the extent to which algorithmic agents are acting in alignment with their higher order preferences, and this must not involve excessive monitoring costs. It would be easy to focus on alignment here and neglect the importance of monitoring costs. This would be a mistake. As individuals, we can improve the quality of our decisions by collecting more information and more thoroughly considering our options. However, since there are costs to gathering and processing information it is not generally desirable to be fully informed and rational on all matters (Gigerenzer & Goldstein, 1996; Taylor, 2020). The same is true for the decisions made on our behalf by agents, whether human or machine. Arrow (1968, p. 538) made this point clearly early in the development of Principal-Agent theory:

> *The principal-agent relation is very pervasive in all economies and especially in modern ones; by definition the agent has been selected for his specialized knowledge and therefore the principal can never hope completely to check the agent's performance. You cannot therefore easily take out insurance against the failure of the agent to perform well. One of the characteristics of a successful economic system is that the relations of trust and confidence between principal and agent are sufficiently strong so that the agent will not cheat even though it may be "rational economic behavior" to do so. The lack of such confidence has certainly been adduced by many writers as one cause of economic backwardness.*

Real-world decision-making requires us to make choices on the basis of imperfect information, and real-world agency relationships require us to empower agents on the basis of imperfect alignment. The institutional challenge, then, involves creating the conditions for meaningful algorithmic choice without imposing an excessive epistemic burden on users. The direct impact is that users could avoid algorithmic agents unlikely to act in accordance with their higher-order preferences. The indirect impact is that users would become more able to assess the trustworthiness of algorithmic agents, thus becoming more willing to employ the trustworthy ones. As in other contexts, this would facilitate exchange by reducing transaction costs (Williamson, 1985).

Such institutions will not completely eliminate the problems caused by type 3 opacity, and the improvements we can expect will not be instant. For one, rational decision-making in a particular context is a capacity developed over time through experience (Smith, 2008). Moreover, institutions supporting such learning need to be created. This reflects a more general pattern in the history of technology: the full impact of a technology is not instant, since complementary innovations (technological, organisational, and institutional) are required. Electrification only transformed manufacturing once factories could be reorganised to fully exploit it benefits, and

information technologies only improved productivity once workers developed the skills to use them effectively and companies found ways to make use of the new tools (Brynjolfsson et al., 2021). Similarly, the capability-enhancing impacts of ADM will only be realised to the extent that institutional innovations are made to enable meaningful choice among algorithmic alternatives.

Acknowledgements We thank the editor and two anonymous reviewers for their insightful and constructive comments and suggestions.

Author Contributions KD and BT jointly conceived the project, conducted the analysis, and prepared the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

Andeweg, R. B. (2000). Ministers as double agents? The delegation process between cabinet and ministers. *European Journal of Political Research*, *37*(3), 377–395. https://doi.org/10.1023/A:1007081222891

Arrow, K. J. (1968). The Economics of Moral Hazard: Further comment. *The American Economic Review*, *58*(3), 537–539.

Beales, H., Craswell, R., & Salop, S. C. (1981). Efficient regulation of Consumer Information, the. *Journal of Law & Economics*, *24*, 491. https://doi.org/10.1086/466997

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 39. https://doi.org/10.3389/fdata.2021.688969

Bergemann, D., & Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, *11*(1), 85–107. https://doi.org/10.1146/annurev-economics-080315-015439

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013.

Black, E., & Hattingh, M. (2020). Assistive Technology for ADHD: A Systematic Literature Review. In T.-C. Huang, T.-T. Wu, J. Barroso, F. E. Sandnes, P. Martins, & Y.-M. Huang (Eds.), *Innovative Technologies and Learning* (pp. 514–523). Springer International Publishing. https://doi.org/10.1007/978-3-030-63885-6_56

Borch, C. (2022). Machine learning, knowledge risk, and principal-agent problems in automated trading. *Technology in Society*, *68*, 101852. https://doi.org/10.1016/j.techsoc.2021.101852

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brennan, G., & Hamlin, A. (2008). Revisionist public choice theory. *New Political Economy*, *13*(1), 77–88. https://doi.org/10.1080/13563460701859744

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, *8*(1), 2053951720983865. https://doi.org/10.1177/2053951720983865

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The Productivity J-Curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, *13*(1), 333–372. https://doi.org/10.1257/mac.20180386

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline&ref=akusion-ci-shi-dai-bizinesumedeia

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Cairney, P., Heikkila, T., & Wood, M. (2019). *Making policy in a complex world*. Cambridge University Press. https://www.cambridge.org/core/elements/making-policy-in-a-complex-world/AACCCA55FEAEFBA971EE261BCAF38575

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 10–23. https://doi.org/10.1093/analys/58.1.7

Coase, R. H. (1937). The nature of the firm. *Economica*, *4*(16), 386–405. https://doi.org/10.1111/j.1468-0335.1937.tb00002.x

Demsetz, H. (1969). Information and efficiency: Another viewpoint. *Journal of Law and Economics*, *12*(1), 1–22. https://doi.org/10.1086/466657

Dowding, K., & Miller, C. (2019). On prediction in Political Science. *European Journal of Political Research*, *58*(3), 1001–1018. https://doi.org/10.1111/1475-6765.12319

Dowding, K., & Taylor, B. R. (2020). *Economic perspectives on government*. Springer International Publishing. https://doi.org/10.1007/978-3-030-19707-0

Durlauf, S. N. (2012). Complexity, economics, and public policy. *Politics Philosophy & Economics*, *11*(1), 45–75. https://doi.org/10.1177/1470594X11434625

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Gaddis, S. M. (Ed.). (2018). *Audit studies: Behind the scenes with Theory, Method, and nuance*. Springer International Publishing. https://doi.org/10.1007/978-3-319-71153-9

Gerdon, F., Bach, R. L., Kern, C., & Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society*, *9*(1), 20539517221089304. https://doi.org/10.1177/20539517221089305

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Hamlin, A. (2006). Political dispositions and dispositional politics. In G. Eusepi, & A. Hamlin (Eds.), *Beyond Conventional Economics: The limits of rational behaviour in political decision making* (pp. 3–16). Edward Elgar.

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., Berking, C., Haferkamp, S., Klode, J., & Schadendorf, D. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, *120*, 114–121. https://doi.org/10.1016/j.ejca.2019.07.019

Hindmoor, A., & Taylor, B. (2015). *Rational choice* (2nd ed.). Palgrave Macmillan.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *ArXiv Preprint ArXiv:1812.04608*.

Ivanov, S. H. (2023). Automated decision-making. *Foresight*, *25*(1), 4–19. https://doi.org/10.1108/FS-09-2021-0183

Jin, Y., & Sendhoff, B. (2008). Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, *38*(3), 397–415.

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

Johnson, M., Albizri, A., Harfouche, A., & Fosso-Wamba, S. (2022). Integrating human knowledge into artificial intelligence for complex and ill-structured problems: Informed artificial intelligence. *International Journal of Information Management*, *64*, 102479. https://doi.org/10.1016/j.ijinfomgt.2022.102479

Kästner, L., & Crook, B. (2023, November 3). *Explaining AI Through Mechanistic Interpretability* [Preprint]. https://philsci-archive.pitt.edu/22747/

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, 103459. https://doi.org/10.1016/j.artint.2021.103459

Kim, E. S. (2020). Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective. *Technology in Society*, *63*, 101378. https://doi.org/10.1016/j.techsoc.2020.101378

Klein, D. B. (2002). The demand for and supply of assurance. In T. Cowen, & E. Crampton (Eds.), *Market failure or success: The New Debate* (pp. 172–192). Edward Elgar.

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta Psychologica*, *213*, 103250. https://doi.org/10.1016/j.actpsy.2020.103250

Kochenderfer, M. J., Wheeler, T. A., & Wray, K. H. (2022). *Algorithms for decision making*. MIT Press.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Krafft, P. M., Young, M., Katell, M., Lee, J. E., Narayan, S., Epstein, M., Dailey, D., Herman, B., Tam, A., Guetler, V., Bintz, C., Raz, D., Jobe, P. O., Putz, F., Robick, B., & Barghouti, B. (2021). An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 772–781. https://doi.org/10.1145/3442188.3445938

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Marette, S., Bureau, J. C., & Gozlan, E. (2000). Product safety provision and consumers' information. *Australian Economic Papers*, *39*(4), 426–441. https://doi.org/10.1111/1467-8454.00102

Mele, A. R. (1992). Akrasia, Self-Control, and second-order desires. *Noûs*, *26*(3), 281–302. https://doi.org/10.2307/2215955

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, *14*(4), 272–344. https://doi.org/10.1561/1100000083

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT Press.

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, *2*(3), 49. https://doi.org/10.1007/s44206-023-00074-y

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Papantonis, G., & Belle, V. (2023). Model Transparency: Why do we care? *Proceedings of the 15th International Conference on Agents and Artificial Intelligence, 2023*, 650–657. https://doi.org/10.5220/0011726300003393

Pavlou, P. A., & Gefen, D. (2004). Building Effective Online marketplaces with Institution-Based Trust. *Information Systems Research*, *15*(1), 37–59. https://doi.org/10.1287/isre.1040.0015

Popkin, S. L. (1995). Information shortcuts and the reasoning voter. In B. Grofman (Ed.), *Information, participation and choice: An economic theory of democracy in perspective* (pp. 17–35). University of Michigan Press.

Portuese, A. (Ed.). (2022). *Algorithmic Antitrust* (Vol. 12). Springer International Publishing. https://doi.org/10.1007/978-3-030-85859-9

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. https://doi.org/10.1145/3514094.3534181

Rodrik, D., Subramanian, A., & Trebbi, F. (2004). Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth*, *9*(2), 131–165. https://doi.org/10.1023/B:JOEG.0000031425.72248.85

Rose, F. (1999). *The economics, concept, and design of information intermediaries: A theoretic approach*. Physica-Verlag

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Schwartz, B. (2009). *The Paradox of Choice: Why more is less, revised Edition*. Harper Collins.

Sen, A. (1999). *Development as freedom*. Oxford University Press.

Shapiro, S. P. (1987). The social control of impersonal trust. *American Journal of Sociology*, *93*(3), 623–658. https://doi.org/10.1086/228791

Shliefler, A. (1985). A theory of Yardstick Competition. *Rand Journal of Economics*, *16*, 319–327. https://doi.org/10.2307/2555560

Smart, P. (2017). Extended cognition and the internet. *Philosophy & Technology*, *30*(3), 357–390. https://doi.org/10.1007/s13347-016-0250-2

Smith, V. L. (2008). *Rationality in economics: Constructivst and ecological forms*. Cambridge University Press.

Somin, I. (2016). *Democracy and political ignorance: Why smaller government is smarter* (2nd ed.). Stanford University Press.

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, *24*(2), 109–143. https://doi.org/10.1007/s10462-005-4607-7

Spulber, D., & Yoo, C. (2013). Antitrust, the internet, and the economics of networks. In *The oxford handbook of international antitrust economics, volume 1*. Oxford University Press

Stigler, G. J. (1961). The economics of information. *The Journal of Political Economy*, *69*(3), 213–225. https://doi.org/10.1086/258464

Taylor, B. R. (2020). The psychological foundations of rational ignorance: Biased heuristics and decision costs. *Constitutional Political Economy*, *31*(1), 70–88. https://doi.org/10.1007/s10602-019-09292-4

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The Final Edition*. Yale University Press.

Ullmann-Margalit, E., & Morgenbesser, S. (1977). Picking and choosing. *Social Research*, *44*, 757–784.

Vining, A. R., & Weimer, D. L. (1988). Information asymmetry favoring sellers: A policy framework. *Policy Sciences*, *21*(4), 281–303. https://doi.org/10.1007/BF00138305

Viscusi, W. K., Jr, J. E. H., & Sappington, D. E. M. (2018). *Economics of Regulation and Antitrust* (5th ed.). MIT Press.

Wigner, E. P. (1995). The Unreasonable Effectiveness of Mathematics in the Natural Sciences. In J. Mehra (Ed.), *Philosophical Reflections and Syntheses* (pp. 534–549). Springer. https://doi.org/10.1007/978-3-642-78374-6_41

Williamson, O. E. (1985). *The economic institutions of capitalism*. Free Press

Wolfram, S. (2023, February 14). What Is ChatGPT Doing … and Why Does It Work? *Stephen Wolfram Writings*. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.