INSEAD The Business School for the World®

Working Paper

2021/48/STR

(Revised version of 2021/43/STR)

Human-Algorithm Ensembles

Vivek Choudhary
Nanyang Technological University, vivek.choudhary@ntu.edu.sg

Arianna Marchetti
London Business School, amarchetti@london.edu

Yash Raj Shrestha
Strategy Management und Innovation ETH, yshrestha@ethz.ch

Phanish Puranam
INSEAD, Phanish.puranam@insead.edu

An “ensemble” approach to decision making involves aggregating the results from different decision makers solving the same problem (i.e., without specialization). We draw on the literature on ensemble decision making in machine learning-based Artificial Intelligence (AI) as well as among human decision makers to propose conditions under which human-AI ensembles can be useful. We argue that human and AI-based algorithmic decision making can be ensembled even when neither has a clear advantage over the other (in terms of predictive accuracy) at a decision task or its sub-components, and even if neither alone can attain satisfactory accuracy in absolute terms. Many managerial decisions have these attributes, and division of labor between humans and AI algorithms is usually ruled out in such contexts because the conditions for specialization are not met. However, we propose that human-AI ensembling is still a possibility under the conditions we identify.

Keywords: Managerial Decision making; Machine Learning; Human-AI Ensembling

Acknowledgements: We thank Carsten Bergenholtz, Harsh Ketkar, Hyunjin Kim, Helge Klapper, Jose Arrieta, Euc Man Lee, Sanghyun Park, Bibek Paudel, Tianyu He, Bart Vanneste, and Kang Xi for their helpful suggestions. The usual disclaimers apply. Puranam acknowledges the Desmarais Fund at INSEAD for supporting the Organizations and Algorithms Research Program.

1. INTRODUCTION

There is considerable interest today among practitioners and academics in the potential for collaboration between humans and artificial intelligence (henceforth, AI). AI endows computers with "the ability to learn without being explicitly programmed" (Samuel 1959, p. 120). While the core concept is not new, recent algorithmic advances coupled with breakthroughs in processing power and the ever-increasing availability of digital data have made AI viable in ways that were not possible before, resulting in a wave of enthusiastic adoption with various applications spanning research and practice. Organizations have started exploring how to improve managerial performance by employing a combination of humans and AI – rather than either alone – to tackle a variety of problems, and scholars have begun to study the antecedents and consequences of such efforts (Murray et al. 2020, Shrestha et al. 2019).

A characteristic common to prior attempts to combine humans (henceforth, H) and AI to perform a task is the emphasis on the division of labor with specialization, whereby humans and AI perform different (sub-)tasks with different types of output, and their outcomes are then combined. This involves redefining the task division and task allocation between them to exploit their respective advantages, in terms of superior output and/or lower cost of labor (Canetti et al. 2019, Holzinger 2016, Murray et al. 2020, Reuters 2018). Consider, for instance, a hiring task that involves screening the application pool and interviewing selected candidates: AI can be used to automate the first sub-task by screening applicants' CVs and shortlisting candidates, and humans to conduct in-depth interviews of the selected few.

However, division of labor with specialization is not the only possibility for H-AI collaboration. In this paper, we consider an alternative that has relevance when the task involves decision making: a division of labor with *ensembling*. In this configuration, the human and AI tackle the same decision problem, but their different outputs are aggregated in some way (e.g., by averaging for estimation problems; quorum, plurality or unanimity for screening or selection problems) to arrive at a solution. Whereas specialization requires H and AI to perform distinct sub-tasks, ensembled H and AI tackle the same task (Csaszar and Steinberger 2021, p. 38). The concept of ensembling H-AI is orthogonal to the ideas of "augmentation" and "automation" (Raisch and Krakowski 2020). Augmentation occurs when

AI is added to humans to improve performance of a task. Automation implies that algorithms replace (at least some) humans in performing the task. Hence, ensembling can correspond to augmentation if AI is added to a group of humans or to automation if an AI replaces some humans in the group.

A variety of literatures have highlighted the potential for improving decision accuracy through ensembling when decision makers vary in the errors they make on the same task (Nisbet et al. 2009, Page 2010, 2014, Tumer and Ghosh 1996). This diversity in errors can (though need not always) be beneficial for ensembling if the errors cancel out, as in the case where two agents make opposite prediction errors: one underestimates and the other overestimates the outcome value. Their respective errors cancel each other out, resulting in the average predicted value being closer to the true value. However, while these literatures so far have considered ensembles consisting of only humans (henceforth, H-H) (e.g., the wisdom of crowds) or only AI (henceforth, AI-AI) (e.g., bagging models), ensembled decision making involving a combination of humans and artificial intelligence (henceforth H-AI) has not received as much attention.

In this paper, we argue that H-AI has qualitatively distinctive features compared to H-H and AI-AI (which can be treated as a single AI meta-algorithm). The unique value added by humans to H-AI ensembles lies in their hard-to-externalize data – commonly given labels such as expertise, intuition, gut-feeling, judgement or even life experience. In contrast, what AI uniquely adds to the H-AI ensemble is the potential to estimate the best fitting function – of arbitrary complexity – that can describe the data it has access to. This also implies that AI can be tuned to generate the best ensemble with a human, taking into account the latter's characteristics (the equivalent of co-specialization or training together in H-H, but with far more tunability). Further, preserving diversity of predictions may be easier within H-AI than H-H ensembles because of limits to explainability and conformity pressures in the former. Combining these ideas, we propose specific conditions under which H-AI ensembles should outperform H-H and AI-AI ensembles in terms of decision accuracy, when H-AI ensembles will be formed through augmentation vs. automation (replacement), as well as the boundary conditions that make the presence of H in the H-AI ensemble indispensable.

While our arguments are applicable to any decision task whose performance can be assessed through accuracy metrics, we see them as particularly relevant to the management context. Managerial activities involve a significant decision-making component in processes such as resource allocation or candidate selection (Mintzberg 1975), where AI cannot (yet) outperform human decision making because the underlying structure of the decision problem is unknown and not enough data is available to study past behavior and patterns to approximate that structure in an unbiased manner (Bao et al. 2020). A division of labor with specialization is unsuitable to such cases, as algorithms offer no clear advantage by taking over a task or its sub-components from humans.

However, since human managers may also often be inaccurate in their decision making (Csaszar and Steinberger 2021), managerial decision problems may still be suited to ensembling between H and AI. In other words, by aggregating the decisions of H and AI algorithms – even if they individually offer an unsatisfactory solution and neither is superior to the other – the decision can be more accurate than relying on either one. Such ensembles can particularly benefit managerial decision making as even modest improvements in accuracy often lead to significant returns (Agrawal et al. 2018, Athey 2018, Cockburn et al. 2018). Moreover, unlike the division of labor with specialization, ensembling is easier to reverse since the counterfactual is always observable in terms of performance (accuracy in decision making) that H and AI independently achieve. This flexibility may be particularly useful in managerial decision contexts when the task environment is changing rapidly.

The theory of H-AI ensembles we propose thus expands the range of activities at which human and artificial intelligences can collaborate beyond the domain of division of labor with specialization. It also has implications for the conditions under which humans are most at risk of being replaced by algorithms in decision making tasks where predictive accuracy is the only metric of success. An important caveat is that unlike a descriptive theory that explains observed phenomena, our theory is a normative one. It offers predictions based on internally consistent arguments about what should occur (i.e., work better), rather than why we already observe a particular empirical pattern (Santos and Eisenhardt 2005). Specifically, as a design theory, it is forward looking, describing possibilities that have yet to be realized or carefully examined. Following Simon, (1996), we believe that such an approach to theory can contribute to progress in a design-centric science such as organization design, to which our theorizing

contributes by explicating the conditions for successful collaboration between agents – human and artificial – in a system with the objective of making good decisions (Burton and Obel 1984, Mintzberg 1979).

2. PRIOR LITERATURE ON COLLABORATION BETWEEN H AND AI IN DECISION TASKS

In this section, we review the relevant literature on how humans and AI can be combined for decision making. We first highlight the close links between prediction and decision making, and why AI algorithms, despite their considerable power at prediction tasks, do not prove universally superior to human decision makers. Next, we review research pertaining to the gains from specialization of humans and algorithms to sub-tasks they are each superior at, as well as the literature on the use of humans as superior decision makers that can act as gatekeepers or trainers of algorithms. We conclude this section by noting that the possibility of ensembling between H and AI – which involves neither specialization nor human superiority – has so far remained under-explored.

2.1. Decision making as prediction

All decisions under uncertainty ultimately involve prediction (Agrawal et al. 2018). When selecting between alternatives – be it which candidate to hire, which project to invest in, or which course of action to pursue for a given strategy – the decision maker predicts the corresponding outcome of each available option and selects the alternative most likely to yield the best outcome. Decisions may involve predictions in the form of (i) estimation (i.e., deciding on the value of a variable, for instance how much to invest in a new project), or (ii) screening (e.g., accepting or rejecting a proposal, such as hiring a candidate for a particular job). Estimation is also referred to as a regression problem, and screening as a classification problem in the AI literature (Hastie et al. 2001).

Existing research documents that AI algorithms can be helpful in tackling decision problems. In contrast to traditional optimization algorithms such as branch and bound, dynamic programming, and integer programming used in the development of expert systems (a form of "traditional" AI) based on exact predetermined rules to identify a solution (Kanet and Adelsberger 1987, Lee et al. 2010), contemporary AI algorithms do not require as much a priori knowledge of the problem structure. Instead they can approximate it based on model fitting to data. In other words, AI algorithms assume a certain

problem structure and fit and evaluate a set of possible models on the data.[1] The most powerful results in terms of predictive accuracy have been generated by the AI architectures known as "deep learning," i.e., multi-layered artificial neural networks (LeCun et al. 2015).

A well-known set of theorems prove the existence of neural network architectures that can effectively capture arbitrarily complex patterns in a dataset. These are broadly known as universal approximation theorems (henceforth, UAT; for details see Kratsios, 2020). These theorems demonstrate the power of neural networks in approximating (or reverse engineering) any arbitrary function using data generated from that function. Sequential combinations of linear and non-linear components in a neural network guarantee that there exist networks such that for every possible input *x,* the network can produce the value *f(x)* (or its close approximation) irrespective of the nature of the function *f*. Strikingly, such theorems show that for *any given set of data on past decisions and their accuracy, an AI algorithm exists* (specifically, in the form of a neural network architecture) *that can be at least as accurate as any other predictive model for future decisions* (including the ones possibly underlying human cognition) that relies only on this data.

Given such power and universality in function approximation, why do AI algorithms not simply overwhelm and replace humans in terms of making accurate predictions to support decisions? Some of the possible reasons include constraints on computational power to fit a sufficiently complex model within reasonable time, and the desire to retain interpretability in human terms (Babic et al. 2021). [2]

However, as we will elaborate in the next section, even if we were to assume away such constraints, an additional important reason why AI algorithms do not displace human decision makers lies in data limitations. Not all relevant data can be captured and codified in a way that can be used by AI algorithms. Indeed, exactly what the relevant data is for decision making is itself often unclear. For instance, consider a managerial prediction problem consisting of making hiring decisions. Which aspect

[1] Note that this is true for both supervised and unsupervised machine learning, because in both cases the algorithm evaluates a loss function. The main difference is that in supervised AI, the loss function contains the target term (e.g., *Y*, as in mean squared error), while this is not the case for unsupervised AI (e.g., total mean distance in k-means clustering).

[2] The UAT results also do not say anything about whether deep neural network architectures capture how human minds work, which is a consideration if the goal is to understand human cognition rather than maximize predictive accuracy (e.g., Hawkins, 2021).

of the candidate's CV interacts with the manager's life experiences and expertise in shaping how he/she decides on whether to hire them or not? While the information on the CV (e.g., education background, work experience, skills) can be potentially codified into data that an algorithm can process, the manager's life experiences (e.g., some form of intuition that evaluates the candidate with potential job fit) may not. If the latter is more important in determining the accuracy of the final decision, then despite the power of AI to produce the most accurate possible prediction based on the codified information from the CV, it will not outperform – and may even be beaten by – the human in terms of final decision accuracy (i.e., making a good hire).

Such considerations have naturally led to a division of labor between H and AI that involves specialization, where each takes on the (sub-)task at which they perform best. For instance, AI algorithms can make predictions about effective hires based on comparing CVs of successful past hires to CVs of current applicants, and humans can form an assessment based on interviews.

2.2. Human-algorithm collaboration with specialization

In an attempt to systematize the research on human collaboration with AI, Dellermann et al. (2019) offer a conceptualization of hybrid systems as "using the complementary strengths of human intelligence and AI, so that they can perform better than each of the two could separately" (p. 637). They also provide a useful taxonomy highlighting the main design dimensions of hybrid systems, spanning task characteristics and task representation to learning paradigms and types of H-AI interaction. The underlying logic in all these instances is based on division of labor with specialization: each agent performs different, non-overlapping subtasks based on their respective skills and capabilities (Agrawal et al. 2018 section 6), in turn yielding economic benefits related to cost effectiveness, speed of task performance, and expansion of scale and scope (Iansiti and Lakhani 2020).

While different authors have highlighted different aspects of human-algorithm collaboration, there is a common emphasis on the underlying logic of specialization in which humans and algorithms take on tasks they are distinctively better at performing (Jarrahi 2018, Murray et al. 2020, Seeber et al. 2020). For instance, across a range of applications (e.g., automated call centres where language-understanding systems handle incoming queries; military drones that fire at targets based on remote human decisions; facial-recognition systems that help immigration officers identify suspect travelers;

image-recognition algorithms that help doctors diagnose diseases), algorithms take over tasks that they do better or at least in a more cost-effective way and with comparable quality to humans. According to industry reports, these applications can generate outcomes two to more than six times better than those involving humans or algorithms alone on several tasks (Daugherty and Wilson 2018).

Situations in which "business as usual," as well as "unusual circumstances" need to be handled are also suitable for division of labor with specialization. Algorithms can deal with normal decisions ("business as usual") and humans can intervene when a regime change (commonly known as "data shift") or a steep decline in quality of algorithmic decisions is detected (broadly corresponding to "unusual circumstances"). Such a division of labor relies on the assumption that: (i) the change is detectable, and (ii) under the unusual circumstances, the task has changed to one where humans outperform the algorithm. This could occur, for example, during an exogenous shock that results in a drastically changed environment from the one used to train the algorithm. Thus, in a dynamic environment, atypical issues can be escalated for human oversight while AI can handle more typical situations (Attenberg et al. 2015, Kamar 2016). In sum, they work collaboratively on different subtasks where the AI is subordinate to a human supervisor, who intervenes as needed when the AI is not dependable.

2.3. Human superiority as a basis for approval and training of AI decisions

Another stream of literature has focused on identifying methods to incorporate input from humans (who are assumed to be superior decision makers compared to AI algorithms) to improve the algorithm's predictive performance. Studies in this domain (e.g., Jain, Munukutla, & Held, 2019; Vellido, 2019) have covered applications in areas such as health imaging analysis and keyword identification. This research can be broadly categorized in two streams according to the type of H-AI interactions: (i) the human as gatekeeper, and (ii) the human "in the loop." The role of humans in both configurations is to correct the errors of the AI and improve the predictive model with human input.

As a gatekeeper, the human agent checks and approves the outcome from AI to mitigate potential prediction errors. The human role is considered necessary and blind reliance on AI without

human supervision may have negative consequences.[3] Human intervention is viewed as pivotal in these cases, and as gatekeepers, humans have the final say in the decision.

In the human-in-the-loop (henceforth, HIL) configuration, humans are focused primarily on the algorithm training process to improve its accuracy (Holzinger 2016). The human takes the role of trainer and is assumed to be endowed with superior insight that can be used to correct the algorithm. The "active learning" framework in the AI literature, where a learning algorithm improves its predictive accuracy by interactively querying a user (human) to verify its prediction and label new data points with the desired outputs, falls into this category (Settles 2012). Unlike gatekeeping, in an HIL setting the goal is to make the algorithms as capable as humans after being trained.

2.4. Summary

The existing literature on human-algorithm collaborative decision making emphasizes either benefits from specialization (i.e., H and AI are each superior at different subtasks) or the superior ability of humans (e.g., the gatekeeper or HIL configurations). The dangers of replacing humans with AI have also been recognized, which include short-termism and reducing variance within organizations, as noted by Balasubramanian, Ye, & Xu (2020). In contrast, our aim is to describe how to combine human and AI-based algorithmic decision making when neither agent has a clear advantage over the other at a task or its sub-components, and even if neither alone can attain satisfactory accuracy in making predictions that underlie the relevant decisions.

Such conditions characterize many managerial decision-making contexts. Whereas division of labor with specialization or human as gatekeeper/trainer paradigms would logically be ruled out in such cases (possibly leaving such decisions entirely in human hands for reasons of tradition, trust, and legitimacy), we believe that ensembling offers an alternative and overlooked path.

[3] For example, Amazon found that an AI algorithm designed to screen job applicants amplified biases in the training data, resulting in unfair outcomes for female candidates (Reuters 2018). Canetti et al. (2019) provide another example of bias exacerbation in the application of AI algorithms to criminal conviction decisions, where AI assists humans by computing the probability of a person being convicted of a crime. It is now known that the assessment will have shortcomings (e.g., because of biased data) that may create unfair outcomes.

3. THEORY: THE DISTINCTIVE BENEFITS OF H-AI ENSEMBLES

The distinctive feature of ensemble decision making is that all its members perform the same prediction task, i.e., there is no specialization (Anderson 2019, p. 23, Brown 2010, p. 1). Ideas relating to ensembles have been extensively studied in both machine learning (e.g., boosting, bagging, stacking, cross-learning algorithms) and the social sciences (e.g., Condorcet's jury theorem, wisdom of crowd, pooling of experts). The existing literature identifies benefits not only from ensembling estimation tasks (e.g., Page 2007), but also screening decisions (e.g., through voting systems) and probabilities or quantile estimates in various ways (Becker et al. 2021, Lichtendahl et al. 2013, O'Hagan 2006, Ranjan and Gneiting 2010, Thomas and Ross 1980).

However, an attempt to understand how and when the ensembling of humans and AI can be useful in (managerial) decision making is novel. To build the foundations of our argument, we start by considering why diversity in predictions (and therefore prediction errors) is necessary (but not sufficient) for ensembles to improve on their best members. Next, we discuss the sources of diversity in prediction errors, and in particular, what AI and H individually bring to the ensemble. We build theoretical propositions about the conditions under which H-AI ensembles outperform other ensembles, as well as when they are likely to be formed by augmenting vs. replacing humans in existing ensembles. We conclude with boundary conditions that must be met for H-AI to dominate other forms of collaboration in practice, not only in theory.

3.1 Why diversity in predictions is necessary (but not sufficient) for ensembles to be useful

Diversity in predictions (and therefore in prediction errors, i.e., the difference between the predictions made and the actual outcome) made by two agents is necessary but not sufficient to improve the accuracy of their ensembled prediction. The intuition is most simply expressed for the case of a single point prediction, based on a result known as "ambiguity decomposition" (Krogh and Vedelsby 1995). This theorem exists in multiple fields related to statistics, with the version given by Scott Page as the "diversity prediction theorem" being the most accessible (Page 2010), and widely popularized by Surowiecki (2005) as the "wisdom of the crowd." It posits that the error of a crowd's estimate (which is the average of its members' estimates) is systematically lower than the average individual error (where the error is expressed as the square of the difference between estimate and truth) as long as diversity

(i.e., sum of squared difference in estimates) across individuals is positive. The key expression can be written as:

$$\text{Crowd error} = \text{Average individual error} - \text{Diversity} \quad (1)$$

The ensemble (in this case the crowd's prediction) can thus be expected to always perform better than the *average* accuracy of its members in a single estimation task, which represents the right benchmark if one cannot hope to learn which of the ensemble members is likely to be the most accurate (Page 2010). However, we might be interested in knowing whether ensembling can help us improve on the best individual predictor, or how we could improve the ensemble's accuracy. Unfortunately, the identity in (1) neither guarantees that the crowd (i.e., the ensemble) always beats the best (i.e., most accurate) individual, nor that increasing the diversity of crowd predictions will necessarily increase its accuracy. The reason is that the two terms, whose difference determines the crowd accuracy (i.e., average individual accuracy and diversity) are not independent (see Appendix 1 for an illustration). Therefore, the crowd's accuracy cannot be solely determined by its diversity, but also requires a consideration of individual accuracy.

The principle of balancing average individual errors (bias) against diversity is central in the extensive literature on error cancellation with aggregation of predictions and is well documented in the ML literature (see Brown et al. 2005, Ueda and Nakano 1996 on the "bias-variance-covariance decomposition"). For instance, methods have been developed to balance average individual accuracy and diversity by identifying models with negative correlations between their prediction errors (known as "NC learning"), which can lead to an ensemble with higher accuracy than even the best member model (Reeve and Brown 2018). The intuition is not unlike that in the story of the two statisticians who go out hunting, shoot and miss their mark by the same amount in opposite directions, but claim they succeeded (on average). Crucial to this story is that they are both off the mark in different but self-cancelling ways. Appendix 2 gives details of three broad classes of AI-AI ensembling, namely bagging, stacking, and cross-learning, and how each balances individual bias and group diversity to form ensembles.

Group composition that leverages demographic diversity (as a proxy for prediction diversity, see Hong & Page, 2004) is the primary mechanism for ensembling among humans. Because of

differences in life experiences and cognitive capacities, bringing a diverse group of individuals together for making decisions (through rules such as pooling their estimates or voting) can be seen as an attempt to use ensembling to improve on individual decisions through error cancellation. Condorcet's jury theorem also illustrates error cancellation based on aggregation through voting (Condorcet 1785). The theorem emphasizes both individual accuracy and diversity: the probability of getting an incorrect decision through a majority vote diminishes by adding jury members who are each likely to be correct at least better than chance because they are independent (diverse in their beliefs), such that their probabilities of being wrong can be multiplied and taken in the limit to zero.

In sum, despite the variety of ensemble techniques studied in the literature, all ensembles, whether among humans or among algorithms, gain predictive accuracy by creating and aggregating diversity in predictions and managing the trade-off between average bias and diversity. It is worthwhile reiterating that there is no theoretical guarantee that ensembles will always outperform their best members. However, we do know that this outcome is more likely to arise when there is (a) diversity in prediction errors made by different models, (b) each of which is at least a "**weak learner"** i.e., at least marginally better than chance in its predictive accuracy (Dormann et al. 2018, Reeve and Brown 2018, Rougier 2016, Schapire 1990). These are necessary but not sufficient conditions for ensembles to outperform their best members.

Results from empirical studies are encouraging in showing that the conditions that make ensembling advantageous over component models seem to be often, if not always, met. For instance, in a series of experimental set-ups, Armstrong (2001) found that ensembles resulted on average in error reduction by 12.5 percent, with the amount of error reduced ranging from three to 24 percent as compared to individual decisions (also see Mendes-Moreira et al. 2012). Džeroski & Ženko (2004) provide empirical evidence that ensembling performed better than the best individual model in various classification tasks, such as disease diagnosis. The Netflix contest[4] had made the practice widely popular in applied ML. In this field, it is common to try several models and compare the ensemble performance with those of the member models, ultimately picking either the best member or the ensemble based on

[4] https://en.wikipedia.org/wiki/Netflix_Prize [online access: 27/06/2021]

accuracy. This flexibility to easily "reverse" the decision to ensemble is one of its key distinguishing features compared to division of labor with specialization, where the counterfactual (i.e., non-specialized decision making) is typically harder observe after specialization.

3.2 What Humans and AI each bring to an ensemble

Given that diversity in predictions errors across its members plays such an important role in an ensemble's performance, it is useful to understand the two fundamental sources from which it can arise: ***different models*** and ***different data*** (as well as a ***combination of both differences***) (Csaszar and Ostler 2020, also see Simons et al. 1999).

The first source of diversity is the ***model*** used to make the prediction, i.e., the *result of a process that involves converting data into a representation of the prediction problem*. For H, models refer to the mental representation of the prediction problem, which is how the human represents the environment and processes information while making decisions (Csaszar and Laureiro-Martínez 2018, Csaszar and Ostler 2020). Humans learn from experience, using forms of associative learning (Heyes 2018). Biological and cultural factors influence how they extract insight from a given set of data, i.e., how much of it is processed and how patterns are recognized. The models used by individuals to make predictions from the same data may therefore vary systematically along demographic dimensions (Phillips et al. 2006).

In the case of AI, the model refers to the result of the machine learning process – the representation used to summarize the patterns discovered in the data. Its components include the training process, the loss function used and the model architecture in terms of available hyper-parameters. Just as two humans or two AI algorithms may differ in the models they learn from the same data, a human and an AI algorithm may also differ from each other. Despite several decades of research in neuroscience, we are yet to attain a reliable model of how humans learn and make decisions (Baars and Gage 2010, Blum and Blum 2021). AI algorithms in use today differ from the models that humans use to process data to make predictions – to an (as of yet) unquantifiable degree (see also Hawkins 2021).

The second source of diversity in prediction is ***access to data*** by the respective agents. Even when the same underlying model is used, two humans or two algorithms can arrive at diverse predictions if they access different data. If we denote the data that can be codified into a digital format and made

available to AI for training as "Data Type I," then this type of data is by definition also accessible to humans (even though they may not be able to process it as effectively as AI if they are overwhelmed by its scale)[5]. However, "Data Type II" could also exist –information available to humans but not to AI for training. In the hiring context, for instance, Data Type II might take the form of what the candidate said in interviews, observation of body language and facial expressions that cannot easily be coded but can nonetheless be used (perhaps unconsciously) by humans in decision making (Ibrahim et al. 2021). It might also be difficult to provide Data Type II to an AI because of privacy concerns or regulation, even when it is codifiable. The crucial point we wish to make here is the possible existence of Data Type II (separately from Data Type I), and the inability of AI to access it.

Given these two sources of diversity in predictions, a baseline conclusion is that *AI might replace H when the predictive accuracy of AI (drawing on Data Type I) exceeds the predictive accuracy of H (drawing on Data Types 1 & possibly Type 2 as well)*. As we have noted, the Universal Approximation Theorems (UAT's) guarantee the existence of neural network architectures that cannot be beaten in terms of predictive accuracy given access to all available data. This implies that an AI can be built that can beat or equal a human if only Data Type I exists. To be sure, there are practical considerations that might prevent the use of the appropriate neural network architecture (such as limits on processing power or the desire to retain explainability) which may allow humans a role in these situations, and we will discuss those further under boundary conditions to our theory. However, the point we wish to highlight here is the importance of Data Type II, without which (and in the absence of practical constraints of the form noted above), humans cannot outperform AI in terms of predictive accuracy.

3.3 The case for H-AI ensembles

The horse-race between H and AI is however not the most useful one to examine, given that we know an ensemble could improve on its component members when it is made up of (at least) weak learners that are diverse in their prediction errors. We therefore turn to theorize conditions under which the H-

[5] This data could well include codified observable aspects of past human decisions, possibly incorporating their biases. However, this does not alter our argument in any way.

AI ensembles can be superior to H-H and AI-AI (which can be treated as a single AI). We break down the comparison of H-AI to H-H and AI-AI across three scenarios describing the availability of Data Type I and/or Data Type II.

Consider Case I in Table 1, where we assume that there is insufficient codified and AI-accessible data (Data Type I), but Data Type II exists, which is accessible only to humans. By insufficient, we mean that using this data it is not possible for any model or human to make predictions better than random guesses. In this case, AI models will not produce predictions that are better than random guesses – i.e., they are not even "weak learners." Hence, AI adds no value to the ensemble, and therefore H-AI is unlikely to be the best ensemble.

Next, consider case II, where there is sufficient Data Type I but insufficient Data Type II. In this case, the human can only rely on Data Type I to become at least a weak learner. However, in this case, based on the UATs, the AI alone cannot be beaten in terms of predictive accuracy given access to all available data. This implies that humans – including groups in ensemble configurations – can never beat an AI in situations that look like Case II, in terms of predictive accuracy. H-AI cannot be the optimal ensemble in this case either.

However, case III describes a situation where there exist both sufficient AI accessible data (Data Type I) as well as human-only accessible data (Data Type II). It represents the case in which both H and AI can add value to the ensemble since they are both at least weak learners, and their models and data are diverse. We formalize this argument as follows:

Proposition 1*: H-AI ensembles are likely to be superior to AI-AI or H-H ensembles when (a) adequate data is available for AI algorithms to produce at least weak learners (Data Type I) and (b) humans possess adequate non-externalizable and private data (Data Type II) which enables them (perhaps in combination with AI-accessible Data Type I) to act at least as weak learners.*

---Insert Table 1 about here---

An H-AI ensemble has an additional advantage relative to H-H ensembles: tunability. It is possible to tune the AI based on the H's past predictions to create the best possible ensemble (i.e., one which is best able to produce error cancellation, for instance through Negative Correlation learning). Tuning human diversity is also feasible through co-specialization and the formation of transactive

memory for instance, but the flexibility and control that a human designer exercises over AI algorithms offers an easier path to tune the error diversity in an ensemble. Human cognition and learning still remain incompletely understood (Heyes 2018), which is why education is an expensive and uncertain affair, whereas training algorithms on data is a relatively easier prospect.

A noteworthy implication of Proposition 1 is that H-AI ensembles can be useful even when neither H nor AI achieves particularly high levels of decision accuracy on their own. As long as both are at least weak learners (i.e., make predictions more accurate than random chance), the diversity in their predictions arising from differences in the data and models used by H and AI may (but is not guaranteed to) improve on either alone. In contrast, a division of labour with specialization is ruled out if neither H nor AI achieves satisfactorily high levels of decision accuracy at sub-tasks (with the default often being to keep things in the hands of the humans in such cases). H-AI ensembling therefore opens up a set of possibilities for collaboration between humans and AI that remain invisible under the logic of division of labor with specialization.

3.4 H-AI ensembles formed through augmentation vs. replacement

Ensembles need not only form through combination (i.e. putting together a single human and an AI). If we allow for ensembles with more than two component models and use H-AI to now denote any ensemble which includes at least some human and some AI composition, then H-AI ensembles may arise either through augmentation – i.e., the addition of an AI – or automation – i.e., the replacement of a human. The need to always retain a human in the ensemble may arise, for instance, for regulatory reasons or to allay public concerns. Our theoretical framework is also useful to understand the conditions under which we may see either kind of H-AI ensemble arise.

Consider an initial ensemble of two humans: H_1-H_2 (the logic below generalizes to any number of humans, and there is no need to separately consider AI-AI ensembles since they can be treated as a single AI that uses an ensemble algorithm). When does it make sense to replace one of the humans with an AI (automation, leading to H_1-AI or H_2-AI) vs. augment the ensemble with an AI (leading to H_1-H_2-AI)? We know that for an AI to have a role in an ensemble it must at least be a weak learner, which in turn implies there must exist Data Type I (Proposition 1). The existence of Data Type I is thus a necessary condition for H-AI ensembles formed either through augmentation or through automation

(replacement). Rather, the question of interest is when it is sensible to keep both humans in the ensemble vs. displace one of them.

Since the advantage of an ensemble of weak learners arises from the diversity of their prediction errors, we can infer that the more similar the prediction errors of H_1 and H_2 are (because of overlapping Data Type II, similar mental models, or both), the less useful it is to keep both. This leads to:

Proposition 2*: H-AI ensembles are more likely to be formed by augmentation rather than automation (replacement) the more diverse the prediction errors made by humans.*

It is worth highlighting two surprising corollaries of Proposition 2. First, when forming an H-AI ensemble through automation, it is not necessary to replace the less accurate human. Rather, the best combination of H-AI should be retained, which may arise by keeping the less accurate human in the ensemble. This is because what matters is not only bias but also diversity in an ensemble (refer to Appendix 1 & 2 for more details).

Second, it is not the overlap in models or data between human and AI that puts the human at risk of replacement through automation in the construction of an H-AI ensemble. Rather, it is the lack of diversity among the humans themselves. In fact, diversity may be easier to preserve in an H-AI ensemble compared to H-H ensembles. In the case of H-H ensembles, social conformity pressures often lead actors to generate similar decisions (Asch 1956, Janis 1982). This is especially common when the group involved in the decision-making task spans multiple levels in the organizational hierarchy, with subordinates being prone to conforming to the manager's decision. Cognitive bias and homophily can also reduce diversity in human groups (McPherson et al. 2001). On the other hand, factors like algorithm aversion (Dana and Thomas 2006, Dietvorst et al. 2015, 2018, Hastie and Dawes 2009, Logg et al. 2019, Luo et al. 2019, Meehl 1954) and the challenge of explaining algorithmic decisions (Lipton 2018) to a human – that are often blamed for limiting the diffusion of H-AI collaboration within organizations – can usefully preserve diversity of predictions in a H-AI ensemble.

3.3 Boundary conditions

We have assumed that, absent Data Type II, H can contribute little to an ensemble with AI, since at least in theory a neural network architecture exists that can produce the best approximation to the function

that generated the data (Lin and Jegelka 2018, Le Roux and Bengio 2010). However, in practice, finding the appropriate network architecture involves a laborious search through the hyperparameter space, and an alternative is to check if the diversity in predictions produced by a human can help to improve on the imperfectly hyper-parameterized network. Yet this is by no means a stable or "safe" state for humans to protect against displacement, as technologies for tuning network architectures become more rapid, efficient and automatic (He et al. 2021).

For the same reasons, even when Data Type II exists, the role of the human will be unstable in the H-AI ensemble, unless the human remains "emulation-proof." Even without access to Data Type II, but with access to Data Type I and a human's past predictions, an AI algorithm can emulate the human's predictions (effectively "inferring" Data Type II), and thus, making the human redundant. Through emulation, Data Type II is effectively converted into Data Type I. This implies that the human's predictions must be noisy, to make it infeasible for AI to perfectly infer Data Type II from the predictions made. At the same time, the predictions made by H should not be so noisy as to prevent the human from adding any accuracy, namely, the predictions must be sufficiently strongly anchored in Data Type II that H actually contributes some predictive power (i.e., H is at least a weak learner). This implies an optimal level of inaccuracy in human predictions based on Data Type II for humans to retain a role in the H-AI ensemble.

Finally, it is possible that changes in underlying data generation processes (i.e., regime changes) alter the value of available data, of Types I and /or II, or bring into existence such data. If the existence of Data Type I and Type II can change over time, these will naturally lead to a changed evaluation of whether an H-AI ensemble is the best configuration for decision accuracy for a particular task.

4. APPLICATION OF H-AI ENSEMBLING TO MANAGERIAL TASKS

Managerial work often involves decisions (e.g., Mintzberg 1975). Further, it is a domain in which human intuition and judgement is often invoked to justify decisions – suggesting the existence of what we have called Data Type II (Acar and West 2021). We can therefore infer that the managerial decisions for which H-AI ensembles will be fruitful to investigate are those for which there is sufficient Data Type I to train AI at least up to the level of a weak learner.

With the digital revolution, organizations can gather large volumes of data and digitize it in a way that may enhance the stock of Data Type I (Adner et al. 2019). AI can be involved in the decision-making process to a larger extent, since it is possible to use that data to train algorithms. Pertinent examples include collecting fine-grained data about the hiring process, the allocation of capital and resources to established and new projects, and the choice of locations to open new branches and points of sale. These are strategic decisions, made by managers, yet performed often enough that firms can collect and codify large volumes of data (Data Type I) required for AI training. Accordingly, we believe that managerial decision making in such contexts can be performed well by H-AI ensembles.

An important pragmatic consideration however is that greater accuracy is not always valued highly enough. Regardless of the type of prediction problem (i.e., estimation or screening), the threshold of acceptable accuracy is set either by the organization or by individuals within it. A manager might have a different desired level of accuracy for hiring employees than that for resource allocation depending on considerations such as profitability, reversibility, or ethical and legal liability for the decision. Additionally, the desired level of accuracy can vary due to the different costs of omission and commission errors. For instance, in the case of stock-picking decisions, a manager can decide on the upper bounds of the permissible omission and commission errors based on an evaluation of risks and opportunity costs (Csaszar 2012).

This is important because we have reasoned about Propositions 1 and 2 in terms of the ensemble that is likely to produce the most accurate predictions, but if, above a threshold, there is no economic value to improving accuracy, then it is possible that the scope of application of H-AI ensembles can shrink if H or AI alone can produce sufficiently accurate predictions on their own. For instance, problems in operational research such as demand forecasting or pricing are examples of tasks traditionally performed by humans that can now be completely handled by AI, because the volume of Data Type I available to organizations for training purposes is enough to make the algorithm sufficiently accurate, and to make an H-AI ensemble unnecessary even if Data Type II existed.

In contrast, several cases of strategic decisions exist that are rare and idiosyncratic enough to prevent firms from collecting and codifying large volumes of AI accessible data (Data Type I). Accordingly, AI algorithms cannot be adequately trained on how to tackle such prediction tasks.

Illustrative examples include the selection of a target company to acquire or a partner to form an alliance with, whom to hire as a new CEO, which new branding campaign to launch, which new industries or market segments to enter. Decisions of this kind are also not suitable for ensembling H and AI either because of a lack of Data Type I.

In sum, whether a managerial task should be tackled by H and AI in ensemble or left entirely in human hands will also depend on whether human accuracy on its own falls below acceptable levels, and whether enough Data Type I is available to organizations for AI training purposes, yet not as much to make Data Type II redundant, in addition to the conditions we have identified in Propositions 1 and 2.

5. DISCUSSION & CONCLUSION

Mirroring the ever-increasing reliance on human-algorithm collaborations witnessed across various industrial domains, organizational researchers have begun to analyze these in theoretical terms (e.g., Balasubramanian et al. 2020, Lindebaum et al. 2020). Many have focused on identifying task and system characteristics and dimensions under which the human or the algorithmic component may be best suited to complete the task (e.g., Dellermann et al. 2019, Jarrahi 2018, Murray et al. 2020). Others have studied how to optimally include superior human judgement as an input to algorithmic processes, either in the form of gatekeeping, (Canetti et al. 2019, Reuters 2018), or human-in-the-loop configurations, where humans train algorithms (Bhardwaj et al. 2020, Holzinger 2016, Jain et al. 2019).

Yet these do not account for the possibility of ensembling decision making by humans and AI algorithms – that neither involves specialization nor requires a comparative advantage of human over AI or vice versa. Moreover, it can be useful even when neither human nor AI on their own attain satisfactory accuracy in their predictive decisions. Ensembling, because of the observability of the counterfactual and the relatively easy reversibility, offers the luxury of being used only when we feel reasonably confident it will be beneficial and can check that belief continuously.

Our paper contributes to the literature on H-AI collaboration in three ways. First, we identify the precise data availability conditions that are likely to make H-AI ensembling attractive. Distinguishing between machine-accessible data and human-only accessible data, we argue that the Universal Approximation Theorems, with their guarantee that AI architectures exist that cannot be

beaten in terms of predictive accuracy for any given data, imply that human value in ensembles must necessarily rely on unique human-specific and machine inaccessible data. The existence of such human-specific data need not be static and can evolve with every decision. Further, humans so far have the lead in being able to learn from unrelated tasks and transfer their learning to the task at hand, thereby constantly enriching the human-only accessible data. This may change if research in machine learning makes significant progress on transfer learning (Zhuang et al. 2019).

Second, we also explain why under the constraint that a human must always be part of an ensemble, the human-AI ensemble can be composed either by augmentation or by automation (Raisch and Krakowski 2020). Somewhat surprisingly, in this situation humans are most at risk of replacement when they exhibit less diversity relative to each other in terms of prediction errors, and not because they have lower accuracy than the AI. In fact, it is not necessarily the human with the lowest accuracy who is at greatest risk of being replaced in the ensemble, but rather the one who adds the least diversity of prediction error relative to the AI.

Third, we also give boundary conditions for the continued importance of humans in human-AI ensembles: the human must be "emulation-proof." This requires human choices based on idiosyncratic, human-specific data to be optimally inaccurate: not so inaccurate that humans achieve no predictive accuracy beyond chance, but noisy enough to prevent AI algorithms from reverse engineering the data uniquely held by humans. In sum, we might expect that AI will displace humans in decisions in which the human-specific data loses predictive power, while humans will likely continue to prevail in decisions where machine-accessible data are not (yet) useful.

5.1. Future Research Directions

While our focus has been on the benefits of ensembling arising from diversity in predictions, further benefits exist to ensembling, which we do not explore in this paper and may provide fruitful avenues for future research on this topic.

A common feature of both human and AI based on machine learning is the possibility of learning from feedback. A domain that seems to offer potential for further investigation is the use of ensembling in learning-by-doing. It is possible that, after each decision event, both humans and algorithms learn from and adapt to the feedback on actions based on past decisions (i.e., learning by

doing (Argote 2013) or reinforcement learning (Sutton and Barto 2018)). Learning-by-doing is different from the process by which algorithms are trained on existing data, for instance in supervised or unsupervised learning. The distinction is sometimes denoted by the terms "online" (i.e., based on actions taken) versus "offline" (i.e., based on pre-existing data) learning (Gavetti and Levinthal 2000, Puranam and Maciejovsky 2020). In online learning, feedback from the task environment (e.g., on how accurate or inaccurate a prediction was) is generated based on past decisions, and agents update their belief about the task based on how accurate their past predictions were, resulting in changes to the underlying prediction model they use.

Ensembling may also create benefits in the online learning-from-feedback process by influencing the diversity of the feedback generated. A single agent that attempts learning-by-doing faces an exploration-exploitation trade-off because of "own action dependence," i.e., they only see feedback on the actions they take, not on the actions that have never been taken (Battigalli et al. 2019). For instance, managers do not observe the performance of employees that they do not hire, though they do observe the performance of stocks they did not invest in. Consequently, in the case of hiring, there is a higher risk that managers will be trapped into hiring a particular type of candidate because of their own priors (exploitation), unless they sample actions inconsistent with their current beliefs (exploration). Diversity in feedback can mitigate this.

Diversity in feedback for different members of an ensemble (even if starting with identical initial beliefs) can arise because of noise in outcomes, or differences in learning processes. The resulting diversity in feedback can be leveraged by exchanging that information among ensemble members. At the same time, the process of sharing experiences may be self-limiting, as it also risks curtailing the benefits of diversity for further learning, thus creating a trade-off between sharing information among ensemble members (allowing them to improve individual performance at any point in time) but at the same time lowering that diversity (Park and Puranam 2020).

If feedback is only available on the ensemble's decisions, then the learning process can be described as learning by participation (Piezunka et al. 2020). In such a process, the feedback obtained is highly dependent on the members of the ensemble who most influence the ensemble prediction. For instance, the final decision made by an ensemble could be determined by voting. In this case, the task

environment only provides feedback on the action proposed by the majority. The potential diversity of feedback on predictions that the less influential members of the ensemble could have contributed is lost. As Piezunka et al. (2020) note, this sets up potential trade-offs between the aggregation rule that allows for best predictions by the ensemble given current data (i.e., how best to exploit existing diversity) versus those that allow for better data to be gathered through feedback on past ensemble decisions (i.e., how to exploit diversity in feedback).

Finally, ensembling may also have some pragmatic benefits in terms of motivation. For instance, humans may experience enhanced epistemic motivation to engage in learning because of the presence of AI as a competitive benchmark. While such a competitive effect may also exist with two humans, the threat of eventual displacement of humans by algorithms and redundancy of humans on the job may serve as a stronger stimulus in the case of human-AI combined learning.

5.2. Conclusion

The possibility of ensembling humans and AI algorithms for decisions that currently neither perform well at, expands the frontiers of human-algorithm collaboration beyond what can be accomplished through specialization or relying on human superiority to act as a gatekeeper or trainer. Since the properties of many important managerial decisions (in particular, their unknown underlying structure and the paucity of data on past decisions) make it difficult for AI to attain satisfactory stand-alone performance in the near term, we suggested that ensembling with managers is a possible avenue worth exploring.

Managers would only become redundant in a division of labor with ensembling if algorithms are able to emulate their decisions, i.e., not just match or even exceed their accuracy, but also make the same errors, so that the human component of the ensemble would cease to contribute any diversity in terms of predictions. To the extent that predicting human behavior remains challenging for both algorithms and fellow humans, this also offers a sanguine picture of the ongoing importance of human managers in an increasingly algorithmic workplace. As long as managers can add diversity of predictions to an ensemble that involves them and algorithms (even if the algorithm meets or exceeds their performance levels), managerial inputs will remain crucial. Even the errors that managers (and humans in general) make, if different, are useful, and not just their insights.

REFERENCES

Acar OA, West D (2021) When an Educated Guess Beats Data Analysis. *Harv. Bus. Rev.*

Adner R, Puranam P, Zhu F (2019) What Is Different About Digital Strategy? From Quantitative to Qualitative Change. *Strateg. Sci.* 4(4):253–261.

Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Review Press, Boston, MA, USA).

Anderson B (2019) *Pattern Recognition: An introduction* (ED-Tech Press, Essex, UK).

Argote L (2013) Organization learning: A theoretical framework. *Organ. Learn.* (Springer), 31–56.

Armstrong JS (2001) Combining Forecasts. Armstrong JS, ed. International {Series} in {Operations} {Research} & {Management} {Science}. (Springer US, Boston, MA), 417–439.

Asch SE (1956) Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol. Monogr. Gen. Appl.* 70(9):1–70.

Athey S (2018) The Impact of Machine Learning on Economics. *Econ. Artif. Intell. An Agenda*. (University of Chicago Press), 507–547.

Attenberg J, Ipeirotis P, Provost F (2015) Beat the Machine. *J. Data Inf. Qual.* 6(1):1–17.

Baars BJ, Gage NM (2010) *Cognition, brain, and consciousness: Introduction to cognitive neuroscience* (Academic Press).

Babic B, Gerke S, Evgeniou T, Cohen IG (2021) Beware explanations from AI in health care. *Science (80-.).* 373(6552):284–286.

Balasubramanian N, Ye Y, Xu M (2020) Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. *Acad. Manag. Rev.*

Bao X, Diabat A, Zheng Z (2020) An ambiguous manager's disruption decisions with insufficient data in recovery phase. *Int. J. Prod. Econ.* 221.

Battigalli P, Francetich A, Lanzani G, Marinacci M (2019) Learning and self-confirming long-run biases. *J. Econ. Theory* 183:740–785.

Becker J, Guilbeault D, Smith N (2021) The Crowd Classification Problem: Social Dynamics of Binary Choice Accuracy. *arXiv Prepr. arXiv2104.11300*.

Bhardwaj A, Yang J, Cudré-Mauroux P (2020) A Human-AI Loop Approach for Joint Keyword Discovery and Expectation Estimation in Micropost Event Detection. *Proc. AAAI Conf. Artif. Intell.* 34(03):2451–2458.

Blum M, Blum L (2021) A Theoretical Computer Science Perspective on Consciousness. *J. Artif. Intell. Conscious.* 08(01):1–42.

Brown G (2010) Ensemble Learning. *Encycl. Mach. Learn.* 312.

Brown G, Wyatt J, Harris R, Yao X (2005) Diversity creation methods: {A} survey and categorisation. *Inf. Fusion* 6(1):5–20.

Burton RM, Obel B (1984) *Designing efficient organizations: Modelling and experimentation* (North Holland).

Canetti R, Ramnarayan G, Cohen A, Scheffler S, Dikkala N, Smith A (2019) From soft classifiers to hard decisions: How fair can we be? *Proc. 2019 Conf. Fairness, Accountability, Transpar.*:309–318.

Cockburn I, Henderson R, Stern S (2018) *The Impact of Artificial Intelligence on Innovation* (Cambridge, MA).

Condorcet M De (1785) *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix* (Imprimerie Royale, Paris).

Csaszar F, Steinberger T (2021) Organizations as Artificial Intelligences: The Use of Artificial Intelligence Analogies in Organization Theory. *Acad. Manag. Ann.*:annals.2020.0192.

Csaszar FA (2012) Organizational structure as a determinant of performance: Evidence from mutual funds. *Strateg. Manag. J.* 33(6):611–632.

Csaszar FA, Laureiro-Martínez D (2018) Individual and Organizational Antecedents of Strategic Foresight: A Representational Approach. *Strateg. Sci.* 3(3):513–532.

Csaszar FA, Ostler J (2020) A Contingency Theory of Representational Complexity in Organizations. *Organ. Sci.* 31(5):1198–1219.

Dana J, Thomas R (2006) In defense of clinical judgment … and mechanical prediction. *J. Behav. Decis. Mak.* 19(5):413–428.

Daugherty PR, Wilson HJ (2018) Humans Plus Robots: Why the Two Are Better Than Either One Alone. *Knowl. Whart.*:1–6.

Dellermann D, Calma A, Lipusch N, Weber T, Weigel S, Ebel P (2019) The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. *Proc. 52nd Hawaii Int. Conf. Syst. Sci.*

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: {People} erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144(1):114–126.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: {People} will use imperfect algorithms if they can (even slightly) modify them. *Manage. Sci.* 64(3):1155–1170.

Dormann CF, Calabrese JM, Guillera-Arroita G, Matechou E, Bahn V, Bartoń K, Beale CM, et al. (2018) Model averaging in ecology: a review of {Bayesian}, information-theoretic, and tactical approaches for predictive inference. *Ecol. Monogr.* 88(4):485–504.

Džeroski S, Ženko B (2004) Is Combining Classifiers with Stacking Better than Selecting the Best One? *Mach. Learn.* 54(3):255–273.

Gavetti G, Levinthal D (2000) Looking Forward and Looking Backward: Cognitive and Experiential Search. *Adm. Sci. Q.* 45(1):113.

Hastie R, Dawes R (2009) *Rational choice in an uncertain world: The psychology of judgment and decision making* (SAGE Publications Inc, CL, USA).

Hastie T, Friedman JH, Tibshirani R (2001) *The Elements of Statistical Learning* (Springer series in statistics New York).

Hawkins J (2021) *A Thousand Brains: A New Theory of Intelligence* (Basic Books).

He X, Zhao K, Chu X (2021) AutoML: A survey of the state-of-the-art. *Knowledge-Based Syst.* 212:106622.

Heyes C (2018) *Cognitive Gadgets* (Harvard University Press, CL, USA).

Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3(2):119–131.

Hong L, Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci.* 101(46):16385–16389.

Iansiti M, Lakhani KR (2020) Competing in the Age of AI. *Harv. Bus. Rev.* 98(1):60–67.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Manage. Sci.* 67(4):2314–2335.

Jain S, Munukutla S, Held D (2019) Few-Shot Point Cloud Region Annotation with Human in the Loop. *ICML Work.*

Janis IL (1982) Groupthink: Psychological studies of policy decisions and fiascoes.

Jarrahi MH (2018) Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Bus. Horiz.* 61(4):577–586.

Kamar E (2016) Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. *Proc. Twenty-Fifth Int. Jt. Conf. Artif. Intell.* IJCAI'16. (AAAI Press), 4070–4073.

Kanet JJ, Adelsberger HH (1987) Expert systems in production scheduling. *Eur. J. Oper. Res.* 29(1):51–59.

Kratsios A (2021) The Universal Approximation Property. *Ann. Math. Artif. Intell.* 89(5–6):435–469.

Krogh A, Vedelsby J (1995) Neural Network Ensembles, Cross Validation, and Active Learning. *Adv. Neural Inf. Process. Syst. 7*:231–238.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.

Lee WC, Chen S kang, Wu CC (2010) Branch-and-bound and simulated annealing algorithms for a two-agent scheduling problem. *Expert Syst. Appl.* 37(9):6594–6601.

Lichtendahl KC, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? *Manage. Sci.* 59(7):1594–1611.

Lin H, Jegelka S (2018) ResNet with one-neuron hidden layers is a Universal Approximator.

Lindebaum D, Vesa M, Den Hond F (2020) Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Acad. Manag. Rev.* 45(1):247–263.

Lipton ZC (2018) The mythos of model interpretability. *Queue* 16(3):31–57.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151(2019):90–103.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: {Machines} vs. humans: {The} impact of artificial intelligence chatbot disclosure on customer purchases. *Mark. Sci.* 38(6):937–947.
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.* 27(1):415–444.
Meehl PE (1954) *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* (University of Minnesota Press, Minneapolis).
Mendes-Moreira J, Soares C, Jorge AM, De Sousa JF (2012) *Ensemble approaches for regression: A survey*
Mintzberg H (1975) The Manager's Job: Folklore and Fact. *Harv. Bus. Rev.* July/Augus.
Mintzberg H (1979) An Emerging Strategy of "Direct" Research. *Adm. Sci. Q.* 24(4):582.
Murray A, Rhymer J, Sirmon DG (2020) Humans and Technology: Forms of Conjoined Agency in Organizations. *Acad. Manag. Rev.* (February):0–44.
Nisbet R, Elder J, Miner G (2009) Model Complexity (and How Ensembles Help). *Handb. Stat. Anal. Data Min. Appl.*:707–721.
O'Hagan A ed. (2006) *Uncertain judgements: eliciting experts' probabilities* (John Wiley & Sons, London ; Hoboken, NJ).
Page SE (2007) Making the Difference: Applying a Logic of Diversity. *Acad. Manag. Perspect.* 21(4):6–20.
Page SE (2010) *Diversity and complexity* (Princeton University Press).
Page SE (2014) Where diversity comes from and why it matters? *Eur. J. Soc. Psychol.* 44(4):267–279.
Park S, Puranam P (2020) Learning what they think vs. learning what they do: {The} micro-foundations of vicarious learning. *arXiv2007.15264 [cs, econ]*.
Phillips KW, Northcraft GB, Neale MA (2006) Surface-Level Diversity and Decision-Making in Groups: When Does Deep-Level Similarity Help? *Gr. Process. Intergr. Relations* 9(4):467–482.
Piezunka H, Aggarwal VA, Posen HE (2020) Learning-by-Participating: The Dual Role of Structure in Aggregating Information and Shaping Learning. *Work. Pap.*
Puranam P, Maciejovsky B (2020) Organizational Structure and Organizational Learning. Argote L, Levine JM, eds. *Oxford Handb. Gr. Organ. Learn.* (Oxford University Press), 520–534.
Raisch S, Krakowski S (2020) Artificial Intelligence and Management: The Automation-Augmentation Paradox. *Acad. Manag. Rev.*:1–48.
Ranjan R, Gneiting T (2010) Combining probability forecasts. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 72(1):71–91.
Reeve HW, Brown G (2018) Diversity and degrees of freedom in regression ensembles. *Neurocomputing* 298:55–68.
Reuters (2018) Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved shorturl.at/dkvNU.
Rougier J (2016) Ensemble averaging and mean squared error. *J. Clim.* 29(24):8865–8870.
Le Roux N, Bengio Y (2010) Deep Belief Networks Are Compact Universal Approximators. *Neural Comput.* 22(8):2192–2207.
Samuel AL (1959) Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* 3(3):210–229.
Santos FM, Eisenhardt KM (2005) Organizational Boundaries and Theories of Organization. *Organ. Sci.* 16(5):491–508.
Schapire RE (1990) The strength of weak learnability. *Mach. Learn.* 5(2):197–227.
Seeber I, Bittner E, Briggs RO, de Vreede T, de Vreede GJ, Elkins A, Maier R, et al. (2020) Machines as teammates: A research agenda on AI in team collaboration. *Inf. Manag.* 57(2):103174.
Settles B (2012) *Active Learning* (Morgan & Claypool Publishers).
Shrestha YR, Ben-Menahem SM, von Krogh G (2019) Organizational Decision-Making Structures in the Age of Artificial Intelligence. *Calif. Manage. Rev.* 61(4):66–83.
Simon HA (1996) The Science of Design: Creating the Artificial. *Sci. Artif.* 111–138.
Simons T, Pelled LH, Smith KA (1999) Making Use of Difference: Diversity, Debate, and Decision Comprehensiveness in Top Management Teams. *Acad. Manag. J.* 42(6):662–673.
Surowiecki J (2005) *The Wisdom of Crowds* (Anchor).
Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (A Bradford Book, Cambridge, MA, USA).

Thomas EAC, Ross BH (1980) On appropriate procedures for combining probability distributions within the same family. *J. Math. Psychol.* 21(2):136–152.

Tumer K, Ghosh J (1996) Error Correlation and Error Reduction in Ensemble Classifiers. *Conn. Sci.* 8(3–4):385–404.

Ueda N, Nakano R (1996) Generalization error of ensemble estimators. *Proc. Int. Conf. Neural Networks*. 90–95.

Vellido A (2019) The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.*

Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2019) A Comprehensive Survey on Transfer Learning.

TABLES

Table 1. Alternative scenarios concerning availability of Data Type I and II, to compare H-AI to H-H and AI-AI ensembles.

	Case I	**Case II**	**Case III**
Adequate Data Type I	No	Yes	Yes
Adequate Data Type II	Yes	No	Yes
Is H-AI likely to be the best ensemble?	No	No	Yes
Examples	New product sales forecast	Dynamic pricing	Recruitment

APPENDIX 1: ILLUSTRATION OF THE BIAS-DIVERSITY TRADE-OFF

Table A1.1 illustrates the trade-off between average individual bias and diversity with an example. Consider a case (A) where two agents solve a prediction problem consisting of estimating the true value of an outcome variable, i.e., the length of a pole in meters, and make predictions of 5m and 6m respectively. If the true value of the pole length is 5m, then the first agent beats the crowd estimate, which predicts 5.5m. Now, consider case (B), in which the second agent is replaced, and the new predictions made by the individuals are of 8 and 6 meters, respectively. These predictions are more diverse than those made by the agents in case (A), corresponding to a gain in crowd diversity from 0.25 to 1. However, the crowd prediction of 7m (the average of the predictions of 8m and 6m) is less accurate than the earlier prediction of 5.5 (the average of 5m and 6m). This example highlights that diversity alone is neither sufficient to improve ensemble accuracy nor does it guarantee beating the best predictor.

Table A1.1. Calculation using diversity-prediction theorem

Case	Individual predictions	Crowd prediction (average)	True Value	Crowd error	Average individual bias	Diversity	Does the crowd beat the best individual?
A	**5.00,** 6.00	5.50	5.00	0.25	0.50	0.25	No
B	8.00, **6.00**	7.00	5.00	4.00	5.00	1.00	No

Note: The best predictor is highlighted in **bold**.

APPENDIX 2: HOW AI-AI ENSEMBLES HANDLE THE BIAS-DIVERSITY TRADE-OFF

Computer scientists have noted that ensemble algorithms can be divided into three broad classes based on the increasing degree to which the ensemble members' diversity is recognized and balanced with average bias, namely: (a) baseline ensembling (or bagging), (b) stacking, and (c) cross-learning. These forms of ensembling differ along two dimensions. First, they differ in whether diversity in prediction stems from the model, the accessible data, or both. Second, they vary in how the individual models are weighted and accounted for in the final ensemble, based on whether such weights are fixed, tuned ex post, or tuned simultaneously to the training step.

In a *baseline ensemble* (or "bagging"), each model is trained independently on a different bootstrapped sample of the same dataset, and the predictions are combined either by averaging or voting over class labels. Bagging can be homogenous or heterogenous in nature, based on similarity or dissimilarity of the member models used. Random forest is an example of homogenous bagging, while bagging between support vector machines and decision trees is heterogeneous. In aggregation, each model is given equal weight. In this class of ensembling, diversity stems from different data access (for homogeneous bagging) and/or different models (for heterogeneous bagging), and there is no weight tuning across the various models.

In *stacking*, a variety of base models is trained independently (similar to bagging), and their predictions are aggregated to build a second meta model that learns how to best combine the base models by identifying the optimal weight for each model. The meta model works like expertise recognition in organisations. First, experts on various tasks are identified such that, during aggregation, the predictions made by the corresponding experts are given different weights. In this second class of ensembling, diversity in prediction stems from both different models and data accessed by the algorithms, and the weights attributed to each model are tuned after the training stage.

Baseline ensemble and stacking approaches to ensemble incorporate stochastic elements during the model construction in the hope of a group of "diverse" predictors emerging with diversity in prediction that enable accuracy gains. In contrast, *cross-learning* approaches (such as boosting and NC-learning) are more direct and explicitly enforce a measure of error diversity on the models. Each model is built to ensure that it is substantially different from other models in its errors, thereby creating model interdependence. Boosting algorithms, such as Adaboost, accomplish this by re-weighting the training examples for each model, increasing the likelihood of more accurate predictions where previous models made more errors. NC-learning takes an even more direct approach by adding a diversity penalty to the loss function, thus managing the accuracy-diversity trade-off while training the member models. Here, diversity stems, as in stacking, from both different models and data access, where the models themselves are tuned during the training phase.

Table A2.1 summarizes the three alternative methods for ensembling and highlights how those vary in the extent to which they create and manage diversity.

Table A2.1. Different ensemble algorithms and their characteristics.

Method	**Extent to which diversity is created and managed**
Baseline ensemble, or bagging	**Low** (individual models are optimized to reduce model errors; each model is given the same weight).
Stacking	**Medium** (individual models are optimized to reduce model errors; each model is given optimal different weights).
Cross-learning ensembles (boosting or NC-learning)	**High** (individual models are optimized to reduce ensemble errors; each model is given optimal weight).