Journal of Management
1–34
DOI: 10.1177/01492063231194968

Human-AI Ensembles: When Can They Work?

Vivek Choudhary
Nanyang Technological University[†]
Arianna Marchetti
London Business School[†]
Yash Raj Shrestha
University of Lausanne[†]
Phanish Puranam
INSEAD

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

Keywords: *managerial decision-making; project evaluation; artificial intelligence (AI); algorithm; organization design; ensembling*

Acknowledgments: We thank participants of the 2021 NYU AI in Strategy Workshop and the 2022 SMS Annual Meeting for their feedback on previous versions of the paper. We are also grateful to the Associate Editor, Gokhan Ertug, and two anonymous reviewers for their very helpful comments and guidance throughout the review process. Vivek Choudhary acknowledges the support from MOE, Singapore grant RS12/20. Yash Raj Shrestha acknowledges funding from Swiss National Science Foundation (grant #215542). Phanish Puranam acknowledges funding from The Desmarais Fund at INSEAD for the Organizations & Algorithms project.

[†]Shared first authorship

Corresponding author: Arianna Marchetti, Strategy & Entrepreneurship Area, London Business School, Regent's Park, London NW1 4SA, United Kingdom.

E-mail: amarchetti@london.edu

Introduction

There is considerable interest today in the potential for collaboration between humans and artificial intelligence (henceforth, AI) technologies. Contemporary AI is based on machine learning (henceforth, ML) techniques that allow computers to learn solutions from data rather than be explicitly programmed (Bishop & Nasrabadi, 2007; Goodfellow, Bengio, & Courville, 2016; also refer to Csaszar & Steinberger, 2021 for a review of the literature). While the core concept is not new, recent algorithmic advances coupled with expansion 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 (e.g., Murray, Rhymer, & Sirmon, 2020; Shrestha, Ben-Menahem, & Krogh, 2019).

A characteristic of prior attempts to combine humans (henceforth, H) and AI to perform a task[1] has been the emphasis on division of labor with specialization. Specialization implies that H and AI perform different (sub-)tasks and their distinct outputs are then combined to generate the final output.[2] Division of labor with specialization involves redefining the task division and task allocation between agents 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; Dastin, 2018). Consider, for instance, a hiring task that involves two sub-tasks: screening the application pool and interviewing selected candidates. AI can be used to automate the first sub-task by screening applicants' resumes and shortlisting candidates, and H to conduct in-depth interviews of the selected few (Pessach et al., 2020). A more subtle form of specialization involves H checking the work of AI or training it (on the presumption that H has superior capabilities to do so). In sum, division of labor with specialization arises when an original task can be decomposed into sub-tasks and those can be allocated across actors based on their relative competence. As we will show, the wide variety of H-AI collaboration literature (including work on the "human in the loop" configuration, e.g., Holzinger, 2016; Ostheimer, Chowdhury, & Iqbal, 2021) implicitly assumes a division of labor with specialization.

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 based on a prediction: A division of labor without specialization. The idea of aggregating the predictions of multiple decision makers tackling the same decision problem has a rich heritage in the literatures on human decision-making (Nisbet, Elder, & Miner, 2009; Page, 2010, 2014; Tumer & Ghosh, 1996) and in computer science on aggregating multiple prediction models (e.g., Polikar, 2012; Sagi & Rokach, 2018; Zhang & Ma, 2012). In keeping with the latter, we use the term "ensembling" to denote aggregation of predictions from multiple models applied to the same prediction problem (Sagi & Rokach, 2018).[3] In this form of division of labor, H and AI tackle the same prediction problem, but their different predictions are aggregated in some way (e.g., by averaging for estimation problems; quorum, plurality, or unanimity for classification problems) to arrive at a final output

(Puranam, 2021). Whereas specialization requires H and AI to perform distinct (sub-)tasks, ensembled H and AI tackle the same task (Csaszar & Steinberger, 2021: 20).

The concept of ensembling H-AI is orthogonal to the ideas of "augmentation" and "automation" (Raisch & Krakowski, 2020) and can result from either, as we also show in this paper. Augmentation of a set of human workers occurs when an AI is added to the humans to improve the performance of a task. Automation, on the other hand, implies that algorithms replace (at least some) humans in performing the task. Ensembling can be achieved through augmentation if AI is added to (a group of) humans, or thorough automation if AI replaces (some but not all) humans in the group (refer to Table 1 for an illustration).

The appeal of ensembling is the potential for improving decision accuracy when decision makers vary in the errors they make on the same task (Nisbet et al., 2009; Page, 2010, 2014; Steyvers, Tejeda, Kerrigan, & Smyth, 2022; Tumer & Ghosh, 1996). This diversity in errors can (although need not always) be beneficial when 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 the literature so far has considered ensembles consisting of only H (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 H and AI (henceforth H-AI) has not received as much attention.

In this paper, we argue that H-AI ensembles have qualitatively distinctive features compared to H-H and AI-AI (where the latter can be treated as a single AI meta-algorithm). The unique value added by H to H-AI ensembles lies in their hard-to-externalize data—commonly associated with expertise, intuition, gut-feeling, judgement, and 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 combination produces some unique characteristics of the H-AI ensemble, which we study in this paper.

Table 1
Use of Augmentation and Automation for Different Forms of Division of Labor: The Case of Hiring

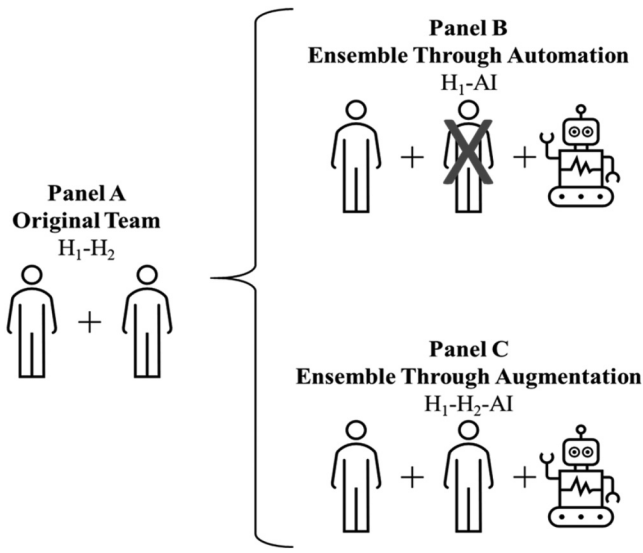
	Tasks Performed by Different Agents Have Same Goal and Outcome [Ensembling, That Is, Non-Specialization]	Tasks Performed by Different Agents Have Different Goals and Outcomes [Specialization]
Augmentation	Newly added AI to a team of two H to make a decision on whether the candidate should be hired for the job; their predictions are aggregated to make a final decision (see Panel C in Figure 1).	Newly added AI decides whether the job candidate reaches a threshold on a set of quantifiable skills that are required for the job. The two H then make the final decision on whether to hire the candidate.
Automation	AI that replaces one of the H (i.e., H_2) and the remaining H (i.e., H_1) both make a decision on whether the candidate should be hired for the job; their predictions are then aggregated to make a final decision (see Panel B in Figure 1).	AI that replaces one of the H (i.e., H_2) decides whether the job candidate reaches a threshold on a set of quantifiable skills that are required for the job. H_1 then makes the final decision on whether to hire the candidate.

While our arguments are applicable to many prediction tasks, in the managerial context they are most relevant to "project evaluation in data-rich contexts." This refers to the task of evaluating a set of alternatives, whose performance can be predicted (and eventually assessed through widely agreed accuracy metrics) on the basis of their features, to ultimately make a choice on which one to select, or how to allocate resources across them (e.g., Christensen & Knudsen, 2010; Csaszar & Ostler, 2020; Sah & Stiglitz, 1985, 1986, 1988). Project evaluation has been used to model the overall decision-making process of managers and firms (e.g., Christensen & Knudsen, 2010; Csaszar & Ostler, 2020), and describes a variety of managerial decision-making scenarios, such as decisions related to hiring employees, investing in projects, selecting suppliers and strategic partners, acquiring firms, or launching new products (Csaszar & Ostler, 2020).

There are many project evaluation problems 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 a sufficiently accurate manner (Bao, Diabat, & Zheng, 2020). A division of labor with specialization may be unsuitable in this case if the AI offers no clear advantage by taking over a (sub-)task from H. However, since human managers might also often be inaccurate in their decision-making (Csaszar & Steinberger, 2021), these problems may still be suited for ensembling between H and AI. In other words, by aggregating the decisions of H and AI—even if AI is inferior to H—the ensembled prediction-based decision can be more accurate than relying on either one alone.

Ensembles can particularly benefit managerial decision-making as, for this class of problems, improvements in decision accuracy often lead to significant economic returns, even

Figure 1
Ensembling Through Automation or Augmentation



with modest improvements in accuracy (Agrawal, Gans, & Goldfarb, 2018; Athey, 2018; Cockburn, Henderson, & Stern, 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.

Unlike a descriptive theory that explains observed phenomena, our theory is a normative one; we cannot yet point to widespread use of ensembles between humans and AI for prediction-based decisions (although there are several instances of the use of AI alone, which we discuss). The recent breakthroughs in generative AI and their rapid adoption across organizations[4] suggest that there may be value in pre-emptive theorizing to anticipate and perhaps guide developments in practice. Therefore, we offer theoretical 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 & Eisenhardt, 2005). Specifically, as a design theory, our paper is forward looking, describing possibilities that have yet to be realized or carefully examined in practice. 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 adds by explicating the conditions for successful collaboration between agents—human and artificial—in a system with the objective of making good decisions based on predictions (Burton & Obel, 1984; Mintzberg, 1979).

Prior Literature on H-AI Collaboration in Decision Tasks

In this section, we review the relevant literature on how H and AI can be combined for prediction-based decision-making. We first highlight the close link 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 H and AI to sub-tasks they are each superior at, as well as the literature on the use of H 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 AI superiority to H in predictive accuracy—has so far remained under-explored.

AI as an Aid to Prediction and Decision-making

All decisions under uncertainty ultimately require prediction (Agrawal et al., 2018), although a prediction may not be enough to make a decision (Bertsimas & Kallus, 2020). For instance, when evaluating projects to select between alternatives—for example, 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, Friedman, & Tibshirani, 2009).

Existing research documents that AI algorithms can be helpful in tackling decision problems involving regression and classification.[5] 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 & Adelsberger, 1987; Lee, Chen, & Wu, 2010), contemporary AI algorithms that are based on ML (which we focus on in this paper) do not require as much a priori knowledge of the problem structure, or clearly defined rules, to generate a solution. Instead, they can approximate it based on models fit to data. In other words, contemporary AI algorithms assume a certain problem structure and fit and evaluate a set of possible models using the data.[6] Such assumptions are defined using a set of model parameters that include, for instance, the degree of interaction between input features, the weights assigned to each input, the non-linearity in the relationships, and so forth. These parameters determine how much each feature will contribute to the final prediction. On the other hand, hyperparameters are top-level parameters that control how the model learns and determine the range of model parameters that the algorithm will estimate (Hastie et al., 2009). Hyperparameters are chosen and set by the analyst before the training of the model even begins, and include, among other factors, the learning rate and stopping rule of the algorithm.

It is this freedom from predetermining exact rules that has allowed the development of complex AI algorithms and driven the recent AI advancements. The most powerful results in terms of predictive accuracy so far have been generated by AI that consists of ML architectures known as "deep neural networks" or "deep learning models," that is, multi-layered complex stacks of artificial neural networks (LeCun, Bengio, & Hinton, 2015).[7] The notable recent advancements in generative AI and related foundation models also rely on deep neural networks trained on large quantities of unlabeled data (e.g., a large part of the internet and public text corpora), at scale and in parallel, using hundreds of computers (Bommasani et al., 2022; Longoni, Fradkin, Cian, & Pennycook, 2022). These deep learning networks specifically rely on advanced "transformer" architectures (Vaswani et al., 2017), and aim to imitate attention-directing mechanisms in human cognitive systems (Shaw, Uszkoreit, & Vaswani, 2018). Our focus in this paper is also on AI algorithms based on neural network architectures, although we exclude generative AI applications for reasons detailed in Section 5.1.

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, UATs; for details see Kratsios, 2021). These theorems demonstrate the power of neural networks in approximating (or reverse-engineering) any arbitrary function to an arbitrary level of precision 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 reproduce the value $f(x)$ (or its close approximation) irrespective of the nature of the function f.

Strikingly, the implication of such theorems is that, given sufficient data on the factors that influenced past decisions and the outcome of the resulting decision, and sufficient computing

power, 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 (including the ones possibly underlying human cognition) to forecast the outcome of future decisions that relies *only* on this data. In other words, for a given data set, the theorems show that there exists a neural network architecture that cannot be beaten in terms of accurately estimating the relationships in that data set if it is large enough and computing power is not a constraint. This result assumes continuous functional forms for the relationships in the data, but attempts are underway to generalize the results, and there already exist many practical solutions using neural network architectures for discontinuous functional forms (Cao, Udhayakumar, Rakkiyappan, Li, & Lu, 2021; Scarselli & Tsoi, 1998).

Given such power in function approximation, why do AI algorithms based on neural network architectures not simply overwhelm and replace H in terms of making accurate data-based predictions to support decisions (barring regulatory constraints)? There are at least two important reasons. First, the existence of the neural network architecture guaranteed by the theorems does not ensure it can be found within reasonable time and cost. For instance, it has taken many years of research to build the complex neural network architectures known as "transformers," as well as the enormous amounts of data and costly computing infrastructures to build Large Language Models like GPT-3 and GPT-4. Training deep learning models is also computationally expensive. For instance, the energy required to train GPT-3 is estimated to be around 1.287 gigawatt hours, which is equivalent to the total electricity consumed by 120 homes in the United States a year (Patterson et al., 2021). Nonetheless, it is true that both data availability and computing power has been increasing dramatically in the recent past.[8] Second, even if we were to assume such constraints away, an additional and perhaps more important reason why AI algorithms do not displace human decision makers lies in limitations in terms of type—not just volume—of data. Not all relevant data can be captured and codified in a way that can be used by AI algorithms (even if there were no data privacy concerns). 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 of the candidate's resume interacts with the hiring manager's life experiences and expertise in shaping how they decide on whether to hire them or not? While the information on the resume (e.g., the candidate's education background, their work experience, and 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 for potential job fit as well as compatibility within the team in a way different from the algorithm) may not. If the latter is more important in determining the accuracy of the final decision, then, despite the power of AI to produce highly accurate predictions based on the codified information from the resume, it will not outperform—and may even be beaten by—H in terms of final decision accuracy (i.e., making a good hire).

Such considerations have naturally so far 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 can make predictions about effective hires based on comparing resumes of successful past hires to resumes of current applicants, and H can form an assessment on a selected pool based on interviews. Many firms now employ deep learning algorithms for resume screening alongside their human teams.[9] In addition to these, there are many other successful implementations of deep learning in managerial tasks, such as expense approval, employee

career feedback, task allocation, and pricing, to name a few. We refer the interested reader to Table A1 in Appendix 1 for a list of successful practical AI applications to managerial tasks.

In sum, similar to any data-driven methodology, contemporary AI has its limitations, such as the requirement of large amounts of data, time, and computational power required for model training, the limited interpretability of outputs by H, the risk of adversarial attacks and catastrophic forgetting, as well as high error rates due to function discontinuity. (We refer the interested reader to Appendix 1 for further details on the limitations of deep learning and current areas for improvements being explored by researchers. Table A3.1 in Appendix 3 also contains a glossary of technical terms used in this section and in the paper more broadly.) Despite these limitations, deep learning-based AI has been effectively used in practice to take on some tasks that it can perform in a manner demonstrably superior to H, or in tasks where its outputs can be easily checked by H (as per the list in Table A1). We next review the literature on these uses, which entail forms of specialization.

Assigning Sub-Tasks to AI

In an attempt to systematize the research on H-AI collaboration, Dellermann and colleagues (2019: 637) offer a conceptualization of socio-technological ensembles (also known as collective or hybrid intelligence) as "using the complementary strengths of human intelligence and AI, so that they can perform better than each of the two could separately." The authors 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 sub-tasks based on their respective 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 & Lakhani, 2020). We refer the interested readers to the rich literature review on hybrid intelligence in Akata et al. (2020).

The common idea underlying the logic of specialization is that H and AI 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 centers 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 suspicious 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 H. According to industry reports, these applications can generate outcomes two to more than six times better than those involving H or algorithms alone on several tasks (Daugherty & Wilson, 2018).

Situations where "business as usual" as well as "unusual circumstances" need to be handled are also suitable for division of labor with specialization. AI can deal with normal decisions ("business as usual") and H can intervene when a regime change (commonly known as "data shift") or a steep decline in the quality of AI 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 unusual circumstances, the task has changed to one where H outperforms the AI. This could occur, for example, due to an exogenous shock

that results in a drastically changed environment from the one used to train the AI. A case in point is online pricing algorithms that failed to effectively predict customer behavior during the COVID-19 pandemic, as the latter rapidly changed their travel and purchase behavior due to unanticipated system shock caused by the pandemic itself (Garg, Shukla, Marla, & Somanchi, 2021). Thus, in a dynamic environment, atypical issues can be escalated for human oversight while AI can handle more typical situations (Attenberg, Ipeirotis, & Provost, 2015; Kamar, 2016).

Approval and Training of AI Decisions by H

Another stream of literature has focused on identifying methods to incorporate input from H (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, 2020) have covered applications in areas such as health imaging analysis and keyword identification. This research can be broadly categorized into two streams according to the type of H-AI interactions: (i) H as gatekeeper, and (ii) H "in the loop." The role of H in both configurations is to correct the AI's errors and improve the predictive model with human input.

As a gatekeeper, the human agent checks and approves the outcome from the AI to mitigate potential prediction errors. The human role is considered necessary, and blind reliance on AI without human supervision may have negative consequences.[10] Human intervention is viewed as pivotal in these cases and, as gatekeepers, H have the final say in the decision. In the human-in-the-loop (henceforth, HITL) configuration, H are focused primarily on the algorithm training process to improve its accuracy (Holzinger, 2016). HITL configurations use complementary strengths to combine H and AI in creating hybrid intelligence configurations (Ostheimer et al., 2021), where H 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 (H) to verify its prediction and label new data points with the desired outputs, falls into this category (Settles, 2012). Unlike gatekeeping, in a HITL setting the goal is to make AI as capable as H after being trained.

Both HITL and gatekeeping are also forms of division of labor with specialization, meaning that the AI and H perform different tasks (e.g., AI makes a prediction and H approves it, or H trains the AI), based on their comparative advantage.

Summary

The existing literature on H-algorithm collaborative decision-making emphasizes the benefits from specialization (i.e., H and AI are each superior at different [sub-]tasks), including the distinction between task performance and training/evaluation (e.g., the gatekeeper or HITL configurations). In applications of the logic of specialization, at least some H are replaced by AI for some sub-tasks (even if no H is made fully redundant). An important risk of this replacement-based approach has been noted by Balasubramanian, Ye, & Xu (2022), who highlight the dangers of short-termism and reduced variance within organizations. Specifically, such a replacement-based approach could also result in automated

human capabilities being unintentionally removed from organizations (e.g., humans losing navigation skills due to the ubiquity of GPS-based navigation).

In contrast, our aim is to describe how to combine H 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 related to project evaluation. Whereas division of labor with specialization (involving assigning some sub-tasks to AI or to H as gatekeeper/trainer) 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.

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, that is, there is no specialization (Anderson, 2019: 23; Brown, 2010: 1). Ideas relating to ensembles have been extensively studied in both ML (e.g., boosting, bagging, stacking, cross-learning algorithms) and the social sciences (e.g., Condorcet's jury theorem, wisdom of the 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, Guilbeault, & Smith, 2021; Lichtendahl, Grushka-Cockayne, & Winkler, 2013; O'Hagan, 2006; Ranjan & Gneiting, 2010; Thomas & Ross, 1980).

However, an attempt to understand how and when ensembling H and AI can be useful in (managerial) decision-making is novel. We build on Steyvers and colleagues' (2022) discussion of the benefits of hybrid systems where H and AI work together for predictive accuracy, as well as on the differentiation of H's and AI's errors in prediction.

We develop our theoretical arguments in three stages. First, we explain why diversity in predictions is necessary but not sufficient for ensembles, rather than just H or just AI decision makers, to be useful. Next, we explain the source of diversity between H and AI predictions. This leads us to identify the precise conditions under which H-AI ensembles will dominate pure H or pure AI decision-making in project evaluation (Proposition 1). Finally, given that an ensemble between H and AI can be formed either by AI replacing one of the existing H (automation) or by adding to the existing H in a team (augmentation), we also derive the conditions under which the former is more likely to add value than the latter (Proposition 2). While we describe our arguments in terms of predictive accuracy, its most direct application in the domain of managerial decision-making is for project evaluation in data rich contexts, such as hiring employees, investing in projects, selecting suppliers and strategic partners, acquiring firms, or launching new products (Christensen & Knudsen, 2010; Csaszar & Ostler, 2020).

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 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 point prediction task, based on a result known as "ambiguity decomposition" (Krogh & 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., the sum of the squared difference in estimates) across individuals is positive. The key expression can be written as follows:

$$\text{Crowd error} = \text{Average individual error} - \text{Diversity} \tag{1}$$

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. Therefore, the crowd's accuracy cannot be solely determined by its diversity, but also requires a consideration of individual accuracy. This limits the usefulness of the wisdom of the crowd approach to our arguments.

Table 2 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—that is, the length of a pole in meters—and make predictions of 5 m and 6 m, respectively. If the true value of the pole length is 5 m, then the first agent beats the crowd estimate, which predicts 5.5 m. Now, consider Case (B), in which the second agent is replaced, and the new predictions made by the individuals are of 8 m and 6 m, 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 one. However, the

Table 2

Trade-off Between Average Individual Bias and Diversity, Calculation Using the 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.

crowd prediction of 7 m (the average of the predictions of 8 m and 6 m) is less accurate than the earlier prediction of 5.5 m (the average of 5 m and 6 m). This example highlights that diversity alone is neither sufficient to improve ensemble accuracy nor does it guarantee beating the best predictor.

To build ensembles that are superior to their best members, the principle of balancing average individual errors (bias) against diversity is central in the extensive literature on error cancellation with aggregation of predictions. This is well documented in the ML literature (see Brown, Wyatt, Harris, & Yao, 2005; Ueda & 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," see Table A3.1 in Appendix 3 for a definition), which can lead to an ensemble with higher accuracy than even the best member model (Reeve & Brown, 2018). Similarly, recent literature shows that diverse errors and Bayesian learning can lead to improved accuracy of a hybrid H-AI predictor in a classification task (Steyvers et al., 2022). 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.

Group composition that leverages demographic diversity (as a proxy for prediction diversity, see Hong & Page, 2004) is the primary mechanism for ensembling among H. Because of differences in life experiences and cognitive capabilities, 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 (i.e., diverse in their beliefs). This allows their probabilities of being wrong to be multiplied and taken in the limit to zero (i.e., attaining accuracy far superior to that of any individual).

In sum, across the variety of ensemble techniques studied in the literature, all ensembles, whether among H or among AI, 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 (a) there is diversity in prediction errors made by different models, (b) each of which is at least a "weak learner," that is, at least marginally better than chance in its predictive accuracy (Dormann et al., 2018; Reeve & Brown, 2018; Rougier, 2016; Schapire, 1990). For instance, to be able to predict whether a coin will show heads or tails (in an experiment on Extra Sensory Perception, for instance), a predictor would be a weak learner if their predictions are accurate more than 50% of the time. Ensemble techniques, in general, try to simultaneously reduce bias and variance of such weak learners by combining several of them together resulting in better performances. When the component models are either identical or do not

perform better than chance, such enhancement in performance via ensembling is unlikely. These are necessary but not sufficient conditions for ensembles to outperform their best members (Dormann et al., 2018; Reeve & Brown, 2018).

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 studies, Armstrong (2001) found that ensembles resulted on average in error reduction by 12.5%, with the amount of error reduced ranging from 3% to 24%, as compared to individual decisions (also see Mendes-Moreira, Soares, Jorge, & Sousa, 2012). Džeroski and 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[11] has 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 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 to observe after specialization has taken place.

What H and AI Each Bring to an Ensemble

Given that diversity in prediction 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 & Ostler, 2020; also see Simons, Pelled, & Smith, 1999).

The first source of diversity is *the model* used to make the prediction by the agents in the ensemble, that is, 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 H represents the environment and processes information while making decisions (Csaszar & Laureiro-Martínez, 2018; Csaszar & Ostler, 2020). H 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, that is, 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, Northcraft, & Neale, 2006).

In the case of AI, the model refers to the result of the ML 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 hyperparameters. Just as two H or two AI algorithms may differ in the models they learn from the same data, a H and an AI algorithm may also differ from each other. Despite considerable progress in research in neuroscience, we are yet to attain a reliable model of how H learn and make decisions (Baars & Gage, 2013), and there is consensus that the AI algorithms in use today differ from the models that H use to process data to make predictions, to an (as of yet) unquantifiable degree (Blum & Blum, 2021; see also Hawkins, 2021).

The second source of diversity in prediction is *access to data* by the respective agents. Even when using the same underlying model, two H or two AI 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 H (even though they may not be able to process it as effectively as AI if they are overwhelmed by its scale).[12] However, "Data Type II" could also exist—that is, information available to H 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, the observation of body language, and facial expressions that cannot easily be coded but can nonetheless be used (perhaps unconsciously) by H in decision-making (Ibrahim, Kim, & Tong, 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. Table 3 summarizes characteristics of Data Type I and II.

Given these two sources of diversity in predictions, a baseline conclusion may be that AI might replace H acting alone in decision-making when the predictive accuracy of AI (drawing on Data Type I alone) exceeds the predictive accuracy of H (drawing on Data Types I & possibly Type II as well). As we have noted, the UATs guarantee the existence of neural network architectures that cannot be beaten in terms of predictive accuracy, given access to all available data and means to protect against overfitting. This implies that, at least in theory, an AI can be built that can beat or equal a H if only Data Type I exists (this is a necessary, not sufficient, condition). 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, among others) which may allow H a role in these situations, and we discuss those further under boundary conditions to our theory in Section 3.5. However, what we aim to highlight is the importance of Data Type II, without which (and in the absence of practical constraints of the form noted above) H cannot be guaranteed to outperform AI in terms of predictive accuracy.

The Case for H-AI Ensembles

The horserace between H and AI is, however, not the most useful one to examine, given that we know that 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-AI ensemble can be superior to H-H and AI-AI (which

Table 3
Characteristics of Data Types I and II

	Data Type I	Data Type II
Accessible by	H & AI	H only
H's ability to process	May be limited, depending on data size	High, possibly sub-consciously
Characteristics	Codified	Tacit & unique to the H's individual experience
Examples (from the context of job candidates hiring)	Information contained in candidates' resumes	What the candidate said in interviews, body language, and facial expressions

can be treated as a single meta-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 Type II.

Consider Case I in Table 4, 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 H. By insufficient data of Type I, we mean that when using this data, it is not possible for any AI or H to make predictions better than random guesses. In this case, AI will not produce predictions that are better than random guesses—that is, 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, H can only rely on Data Type I to become at least a weak learner. However, in this case, based on the UATs, there exists an AI that cannot be beaten in terms of predictive accuracy, given access to all available data. This implies that H—including groups of humans—can be beaten or matched at least in theory by an AI in situations that look like Case II in Table 4, in terms of predictive accuracy. Therefore H-AI cannot be the optimal ensemble in this case either.

However, Case III in Table 4 describes a situation where there is sufficient AI-accessible data (Data Type I) as well as H-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) H 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.

A noteworthy implication of Proposition 1 is that H-AI ensembles can be useful even when neither H nor AI achieves satisfactory levels of decision accuracy on their own, or even if one outperforms the other. As long as both are at least weak learners (i.e., make predictions that are more accurate than random chance), the diversity in their predictions arising from differences in the data and models used by H and AI is likely to (but is not guaranteed to) improve on either alone. In contrast, a division of labor with specialization is ruled out if neither H nor

Table 4

Alternative Scenarios Concerning Availability of Data Type I and Data Type 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

Note. By "adequate data," we mean an amount of data that is needed for an agent to be at least a weak learner (i.e., attain better than chance accuracy) in making a prediction.

AI achieves satisfactorily high levels of predictive accuracy at sub-tasks (with the default often being to keep things in the hands of the H in such cases). H-AI ensembling therefore opens up a set of possibilities for collaboration between H and AI that remain invisible under the logic of division of labor with specialization.

H-AI Ensembles Formed Through Augmentation Versus Automation

If we allow for ensembles with more than two component models—that is, multiple H and multiple AI—and use H-AI to now denote any ensemble which includes at least some H and some AI composition, then H-AI ensembles may arise either through augmentation—that is, the addition of an AI—or automation—that is, the replacement of a H with an AI. 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 H: H_1-H_2 (the logic below generalizes to any number of H, and there is no need to separately consider AI-AI ensembles since they can be treated as a single meta-AI that uses an ensemble algorithm). When does it make sense to replace one of the H with an AI (the case of automation, leading to H_1-AI or H_2-AI) versus augmenting 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 be at least a weak learner, which, in turn, implies that 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 (complementing) or through automation (replacement). The question of interest is when it is sensible to keep both H in the ensemble versus replace one of them with AI.

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 in the ensemble. This leads to the following:

> Proposition 2: H-AI ensembles are more likely to be formed by augmentation (complementing, i.e., adding an AI) rather than by automation (replacing a H with an AI) if the prediction errors made by the H who are already in a team are more diverse with respect to each other. Inter-human diversity increases the value of ensembles formed through augmentation.

We highlight that, in order to compare ensembles formed through augmentation versus automation, one must start with a team initially composed of at least two H. If it were a single H as a starting point, then the choice between automation and augmentation would no longer be a comparison between two types of ensembles. Instead, it would be a choice between a non-ensemble design—that is, either a H or an AI (automation), and an ensemble in which AI augments the single H. Therefore, to consider a fair comparison of ensembles formed through automation or augmentation, we need to begin with a team of at least two H; such a team can then be automated at least partially by replacing one of the H with AI, or augmented by adding an AI to the team of two H. In both cases, the final result is still an ensemble: The difference between the two is that the former has been formed through automation and the latter through augmentation. This then allows for the comparison that underlies Proposition 2.

Figure 1 and Table 1 illustrate how ensembles can be formed through either augmentation or automation. It is worth highlighting two surprising corollaries of Proposition 2. First, when forming a H-AI ensemble through automation, it is not necessary to replace the less accurate H. Rather, the objective should be to retain the best combination of H-AI, which may arise by keeping the less accurate H in the ensemble. This is because what matters for prediction accuracy is not only bias but also diversity in an ensemble (refer to Appendix 2 for more details).

Second, it is not the overlap in models or data between H and AI that puts the H at risk of replacement through automation in the construction of an H-AI ensemble. Rather, it is the lack of diversity among the H themselves. In fact, diversity may be easier to preserve in a 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, Smith-Lovin, & Cook, 2001).

On the other hand, the challenge of explainability of algorithmic decisions (Lipton, 2018) to a H—often blamed for limiting human trust in AI and the adoption of H-AI collaborations within organizations (Glikson & Woolley, 2020)—can usefully preserve the diversity of predictions in a H-AI ensemble. In particular, with the development of deep learning algorithms which could fit arbitrarily complex functional forms, explainability of algorithmic decisions —namely the ability of H to explain why the AI does what it does—currently remains one of the basic challenges of H-AI interactions (Lipton, 2018; Park & Puranam, 2023; Samek, Wiegand, & Müller, 2017). However, a potential inadvertent benefit of this limited explainability might be that it restricts the extent of belief sharing between H and AI, thereby enhancing the preservation of diversity in the ensemble. In other words, H-AI ensembles, where the agents can observe each other's input to and output of the decision-making process, but not the exact function used to generate the prediction, is expected to preserve diversity in prediction to a higher extent than cases where the agents share the same belief system (Park & Puranam, 2023).

Boundary Conditions for the Usefulness of H in Ensembles

We have assumed that, in the absence of Data Type II, H can contribute little to an ensemble with AI since, at least in theory, there is a neural network architecture that can produce the best approximation to the function that generated the data (Le Roux & Bengio, 2010; Lin & Jegelka, 2018). However, in practice, finding the appropriate network architecture involves a laborious search through the hyperparameter space, with the potential alternative consisting of checking if the diversity in predictions produced by a H can help to improve on the imperfectly tuned network. Yet, this is by no means a stable or "safe" state for H to protect against displacement, as technologies for tuning network architectures become more rapid, efficient, and automatic (He, Zhao, & Chu, 2021). "Routine" tasks, that is, tasks that organizations perform repeatedly over time, may be particularly suitable for takeover by AI. This is the case not only because these tasks, given their stable nature, may often be easily codified, but, even if not codifiable, being performed frequently, organizations may produce large volumes of data on them so that AI could easily learn how to optimally perform them.[13]

Further, it is possible that changes over time in underlying data generation processes (i.e., regime changes) alter the value of available Data 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 the H-AI ensemble is the best configuration for decision accuracy, given a particular task. If the value of Data Type II in predictions disappears due to drastic modifications in the data generating process, there would be no benefit in retaining the H component in the ensemble, and the predictive task would be best performed by the AI only, leveraging Data Type I through specialization. However, especially in unstable contexts as in the case of managerial decision-making for project evaluation (e.g., Raisch & Krakowski, 2020), further changes may occur so that Data Type II becomes again available over time, leading to the task being optimally performed with ensembling instead of division of labor via specialization. Yet another possibility is that Data Type I may become irrelevant in the ensemble, and thus the task would be optimally performed by a specialized H that leverages Data Type II.

Application of H-AI Ensembling to Managerial Tasks

Managerial tasks pertaining to project evaluation involve 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 & 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 increasing availability of large volumes of data that organizations can digitize, the stock of existing Data Type I can be enhanced (Adner, Puranam, & Zhu, 2019). Pertinent examples include collecting fine-grained data about the hiring process, the allocation of capital and resources to tasks as well as 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 frequently enough that firms can collect and codify large volumes of data (Data Type I) required for AI training. Accordingly, we believe that such forms of project evaluation are a good candidate to be performed by H-AI ensembles. Note that it is not necessary for the algorithm to attain levels of accuracy comparable or superior to H for it to be valuable in the ensemble. As long as it is a weak learner and adds model diversity, forming a H-AI ensemble may be an improvement on H alone.

An important pragmatic consideration, however, is that greater predictive accuracy attained by the ensemble is not always sufficiently valuable. 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, managers explicitly or implicitly 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 for project evaluation. However, 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 operations research such as demand forecasting or pricing are examples of tasks traditionally performed by H that can now be completely handled by AI (Carbonneau, Laframboise, & Vahidov, 2008), 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 in the form of project evaluation 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 cannot yet be adequately trained to tackle such prediction tasks. Illustrative examples include the selection of a target company to acquire, or a partner to form a merger or an alliance with; whom to hire as a new CEO; which new branding campaign to launch; or which new industries or market segments to enter. Decisions of this kind are also not suitable for ensembling H and AI either, because of the lack of Data Type I, which prevents AI from satisfying the "weak learner" condition. This assessment can change with access to new data sources or new research on AI that can develop algorithms that learn effectively from limited data. For instance, with respect to making accurate predictions about transactions in the Private Equity Industry, Sen and Puranam (2022) showed that, using historical data on past transactions, it was possible to make highly accurate predictions about the type of investors in syndicates formed for particular types of investments. Similarly, recent research in medical AI has developed efficient deep learning approaches for disease diagnosis given high costs of accumulating medical data (Stephen, Sain, Maduh, & Jeong, 2019).

In sum, to implement H-AI ensembles for managerial decision-making of the form of project evaluation, the task needs to first meet the technical conditions set out in Proposition 1, namely that both the H and AI must have access to sufficient data on past instances of the kind of decision they are engaged in solving, to be at least capable of being weak learners. Second, it is important to preserve the independence of predictions of H and AI. This means that, when making their own judgments, H should not be shown the AI's decisions first, and vice versa. Third, a mechanism for aggregating the predictions of H and AI must exist. A practical way to accomplish the aggregation is to have a third individual, for example, a human manager, who takes as input the results of both the H's and AI's predictions and combines them according to an aggregation rule. This means of separating the aggregator from the predictor into two separate human roles can also help to preserve independence between the AI and the H in the prediction stage.

Whether a managerial task concerning project evaluation should be tackled by H and AI in ensemble or left entirely in human hands will depend also on whether improvements in accuracy beyond current levels achieved by H alone is valuable, and whether Data Type I is available to organizations for AI training purposes but not as much to make Data Type II redundant. We assume these boundary conditions are met when stating Propositions 1 and 2.

Discussion and Conclusion

Mirroring the ever-increasing reliance on AI witnessed across various industrial domains, including the recent breakthroughs with generative AI and the ChatGPT applications, organizational researchers have begun to analyze it in theoretical terms (e.g., Balasubramanian et al., 2022; Lindebaum, Vesa, & Hond, 2020). Many have focused on identifying task and system characteristics and dimensions under which the human or 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; Dastin, 2018), or human-in-the-loop configurations, where H train AI (Bhardwaj, Yang, & Cudré-Mauroux, 2020; Holzinger, 2016; Jain et al., 2019).

Yet, these do not account for the possibility of ensembling decision-making by H and AI—which does not involve specialization, or an advantage of H over AI, or vice versa. Moreover, it can be useful even when neither H nor AI on their own attain satisfactory accuracy in their predictions. Ensembling, because of the observability of the counterfactual and the relatively easy reversibility, offers the luxury of being easily discarded when we feel reasonably confident that it will no longer be beneficial, and allows for that belief to be checked continuously. This observability of the counterfactual is particularly useful in dynamic environments due to the chances of AI model decay (i.e., reduction in accuracy) caused by data shifts, where managers can continuously evaluate AI models' accuracy as compared to their own decisions. Additionally, as H are involved in the exact same decision-making, H-AI ensembling also avoids the risk of H losing decision-making capabilities, which is the case with replacement-based approaches (Balasubramanian et al., 2022).

Our paper contributes to the literature on H-AI collaboration in three ways. First, we delineate the possibility for collaboration between H and AI through ensembling, a form of division of labor without specialization. Second, we identify the precise data availability conditions that are likely to make H-AI ensembling attractive to project evaluation tasks performed by managers in data-rich contexts. Distinguishing between AI-accessible data and H-only accessible data, we argue that the human value in ensembles must ultimately rely on unique H-specific and machine inaccessible data. The existence of such H-specific data need not be static and can evolve with every decision. Further, H so far have the lead in the ability to learn from unrelated tasks and transfer their learning to the task at hand, thereby constantly enriching the H-only accessible data. This may change if research in AI makes significant progress on the problem of transfer learning (Zhuang et al., 2020).

Third, we also explain why, under the constraint that H must always be part of an ensemble, the H-AI ensemble can be composed either by augmentation or automation (Raisch & Krakowski, 2020). In automation, a human agent is replaced or substituted by an AI, whereas in augmentation the AI is added to the ensemble to complement the human agents (Tschang & Almirall, 2021). Somewhat surprisingly, in this situation H 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 AI. In fact, it is not necessarily the H 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.

Future Research Directions

While our focus has been on the benefits of H-AI ensembling for project evaluation arising from diversity in predictions, further benefits exist of such ensembles, which we did not explore in this paper and may provide fruitful avenues for future research on this topic.

A common feature of both H and AI based on ML 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 H and AI learn from and adapt to the feedback on actions based on past decisions (i.e., learning by doing, e.g., Argote, 2013). Learning-by-doing reflects reinforcement learning (Sutton & Barto, 2018) and 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 & Levinthal, 2000; Puranam & Maciejovsky, 2020). In online learning, feedback from the task environment (e.g., on how accurate 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 their underlying prediction model. This feedback also enables indirect interactions between the two types of agents.

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 dependence on their own action, that is, they only see feedback on the actions they take, not on the actions that have never been taken (Battigalli, Francetich, Lanzani, & Marinacci, 2019). For instance, managers do not observe the performance of employees that they do not hire, although they do observe the performance of stocks they did not invest in. Consequently, in the case of recruiting, 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.[14] It 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 & Puranam, 2023).

If feedback is only available on the ensemble's decisions, then the learning process can be described as learning by participation (Piezunka, Aggarwal, & Posen, 2022). In such a process, the feedback obtained is highly dependent on the members of the ensemble who influence the ensemble prediction the most. 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, Aggarwal, and Posen (2022) 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).

Ensembling may also have some pragmatic benefits in terms of motivation. For instance, H 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 another H, the threat of eventual displacement of H by algorithms, and the redundancy of H on the job, may serve as a stronger stimulus in the case of H-AI combined learning.

Finally, while in the paper we focus on ensembling for project evaluation in data-rich contexts—that is, a form of prediction tasks that are quantitative in nature—we believe that the H-AI ensembling configuration may also benefit qualitative managerial tasks such as the ones currently solvable by generative AI models—in particular, Large Language Models such as ChatGPT and DALL-E2 (Bommasani et al., 2022; Floridi & Chiriatti, 2020). Generative AI models are considered a turning point in the field of AI and are widely believed to have the potential to revolutionize a wide range of industries and applications, from marketing and sales, operations, human resource management, risk, and legal, to research and development (Mollick & Mollick, 2022).

Given their ability to generate original human-like textual, visual, and auditory content with little or no human input and intervention, drawing entirely on the data on which they have been trained, these technologies are creating many new AI applications. For instance, generative AI can assist in marketing and sales by creating user guides, analyzing customer feedback, identifying potential risks, and improving sales support chatbots (Chui, Roberts, & Yee, 2022). In human resource management, generative AI can be used to create interview questions for candidate assessment and automate first-line interactions such as employee onboarding. The advent of generative AI is currently driving the rapid industrialization of AI, as companies are adopting and customizing existing foundation models into their business processes and products.

However, what these developments mean for H-AI ensembles remains to be seen. In contrast to regression and classification tasks, where the ensembling is relatively straightforward through (weighted) averaging, the aggregation of generative AI outputs with human generative content is much more complicated, due to its qualitative nature. The next step for future research is thus to identify aggregation mechanisms to ensemble H and AI for qualitative outputs. This has potentially significant implications for organizational innovation and creativity. This direction appears promising because many qualitative tasks, such as natural language processing applications that frequently underlie OpenAI solutions, are already solved by ML through a translation into quantitative tasks (e.g., the vectorization of text into embedding spaces).

Limitations

We acknowledge that our theory on H-AI ensembling does not account for all possible interactions between H and AI. First, not all forms of collaboration between H and AI involve prediction-based decisions. For instance, as noted above, our theory does not directly apply to ensembling of qualitative outputs between H and generative AI. Combining the creative outputs (e.g., text, images, music) of H with that of AI is not identical to static error cancellation, but may be closer to the dynamics of learning by doing and coupled search in complex search environments (e.g., Knudsen & Srikanth, 2014; Puranam & Swamy, 2016), which requires further investigations.

Second, we acknowledge that not all managerial decision-making is suitable to H-AI ensembling. While project evaluation has been used to model the overall decision-making process of managers and firms (e.g., Christensen & Knudsen, 2010; Csaszar & Ostler, 2020), and encompasses a wide array of managerial decision-making scenarios, this class of problems does not comprise everything managers do, and excludes, for instance, motivating employees, or complying with legal requirements. Further, the additional condition for a task pertaining to project evaluation to be handled by AI in ensemble with H is that considerable amounts of data on past decisions (e.g., past records of hiring, partner selection, project investment) must exist.

Relatedly, our paper focuses on "project evaluation in data-rich contexts" and does not cover what are known as wicked problems (Grewatsch, Kennedy, & Bansal, 2021; Rittel & Webber, 1973; see Lönngren & van Poeck, 2021 for a comprehensive review of the literature). Such problems (e.g., poverty and environmental degradation) are commonly characterized by high levels of (1) complexity, (2) uncertainty, and (3) divergence in values. In wicked problems, complexity emerges due to the involvement of multiple stakeholders, the existence of various systems, and the interconnectedness of factors, which give rise to intricate interdependencies. In the framework of AI, complex problems can be understood as those fitted by complex functional forms, that is, functional forms with several interaction terms. While deep learning can, in principle, fit predictive models for such complex problems, the real challenge may be data availability.

Given the results of the Universal Approximation Theorems, we know that deep learning algorithms can, in theory, fit predictive models for such complex problems, if there are sufficient data. However, problem uncertainty and divergence among actors produce limits on data. Uncertainty stems from incomplete or missing information, contradictory evidence, and dynamically changing variables. Divergence refers to conflicts in values among different stakeholders (meaning different models and also accessing different sets of data), as they may possess alternate understandings and interpretations of the problem. This divergence complicates the development of a shared outcome as well as a clear evaluation metric for the solution (Camillus, 2008). Therefore, even if deep learning algorithms can assist in addressing the complexity associated with wicked problems, problem uncertainty and the absence of an agreed-upon model and evaluation metric (problem divergence) violates the condition of data availability (of either Type I or II), which is essential to include in a managerial decision-making problem. Wicked problems, however, as currently defined in literature, are beyond the scope of our paper as they cannot be solved by AI algorithms, either standalone or in ensemble, because of the limited availability of data.

Third, while H-AI ensembling allow for direct interactions between agents typical of observational learning (Park & Puranam, 2023), we acknowledge that agents' interdependence in feedback typical of coupled learning (e.g., Knudsen & Srikanth, 2014; Puranam & Swamy, 2016) is beyond the scope of our research. A distinctive feature of combined decision-making by H and AI, as opposed to other forms of technology adoption, is the potential for mutual adjustment: Both H and algorithms not only learn on their tasks from feedback, but they can also learn to adjust to each other and from each other via interaction. The literature on vicarious learning, while focused on interaction among humans (e.g., Levitt & March, 1988; Myers, 2018, 2021), offers a general framework to think about various configurations in which one learning entity can interact with another, and it can be generalized to the case of H-AI ensembling.

With the vicarious learning framework, each learner can have access to the experience of the other, where experience may comprise any combination of past inputs (i.e., attributes of projects), processes (i.e., logic behind action), outputs (i.e., actions), and feedback (i.e., outcomes; Bandura, 1977; Cyert, March et al., 1963). For instance, in a team of one H and one algorithm that produces a recommendation on picking equities to invest in, the H may have access to the attributes, actions, and results produced by the algorithm, and vice versa. Park and Puranam (2023) formally analyze four variants on vicarious learning based on what the agents share with each other: Belief sharing, observational learning (i.e., the case in which attributes, actions, and outcomes are visible), imitation (i.e., the case in which only actions are visible), and inspiration (i.e., the case in which only outcomes are visible).

Our theory of H-AI ensembling allows for some overlap in the data (i.e., experience) that the H and AI access, which implies a form of direct interaction. Specifically, the learning process in this instance resembles what Park and Puranam (2023) define as observational learning, that is, a form of vicarious learning in which the agents learn from the observed attributes, actions, and outcomes of others, but with a relatively low depth of interaction and with limits on what can be observed, which refers to the existence of H-only accessible data (i.e., Data Type II) in ensembles.

We also highlight that, despite H-AI ensembles allowing for direct interactions, in our analysis—and consistent with prior research on vicarious learning—we focus on task environments where the performance feedback on an agent's actions is independent of the actions taken by others (Lazer & Friedman, 2007; March, 1991). An alternative would have been to consider what is known as coupled learning (e.g., Knudsen & Srikanth, 2014; Puranam & Swamy, 2016). Coupling in learning occurs when it is not possible to separate feedback: For instance, when a H and AI jointly produce a recommendation, the feedback would consist in assessing the quality of the aggregate recommendation, rather than of its component parts. Such interdependence in feedback is beyond the scope of our analysis. This is because, when ensembled, each learner produces a complete decision (i.e., ensembles do not entail division of labor with specialization): The feedback can thus be effectively decoupled, since each complete decision can be evaluated on its own.

Conclusion

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

ORCID iDs

Vivek Choudhary https://orcid.org/0000-0003-0374-2723
Arianna Marchetti https://orcid.org/0000-0002-7566-9898
Yash Raj Shrestha https://orcid.org/0000-0002-2699-4723

Notes

1. We refer to a task as a goal-oriented activity to achieve a desired outcome. Different tasks have different outputs that satisfy different goals (von Hippel, 1990; Willis, 1996; Baldwin & Clark, 2000).

2. Specialization is possible because agents possess divergent capabilities ex ante, or will develop them ex post (i.e., divergence in capabilities can occur over time). Throughout the paper, when we refer to "specialization," we mean that actors perform distinct tasks that produce distinct outputs.

3. Ensembling is the opposite of specialization, that is, it occurs when multiple models (or agents) perform the same prediction task (e.g., bagging, stacking).

4. https://www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business (accessed February 2, 2023).
https://www.forbes.com/sites/benjaminlaker/2023/01/06/generative-analysis-how-ai-enables-effective-leadership/ (accessed January 10, 2023).

5. The family of ML models that tackle regression and classification tasks are known as "supervised learning" models.

6. Note that this is true for both supervised and unsupervised ML, 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).

7. Large Language Models such as Chat GPT also involve such deep learning architectures (transformers).

8. https://ourworldindata.org/artificial-intelligence (accessed July 12, 2023).

9. https://cvviz.com/product/resume-screening/ (accessed March 17, 2023).
https://ideal.com/ (accessed November 11, 2022)

10. 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 (Dastin 2018). Canetti et al. (2019) provide another example of bias exacerbation in the application of AI algorithms to criminal conviction decisions, where AI assists H 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.

11. https://en.wikipedia.org/wiki/Netflix_Prize (accessed June 27, 2021).

12. 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.

13. While the specific definition of routine tasks varies in literature (refer to alternate conditions in e.g., Brynjolfsson & Mitchell, 2017, and Acemoglu and Restrepos, 2019), the boundary condition we arrived at is highly consistent with that in extant research: In essence, access to training data that produces at least a weak learner is a fundamental limit on automation through AI. Put differently, if all data accessible to H were also made accessible to AI, then there would be no technical efficiency reasons not to automate the related tasks. There might however be institutional (i.e., legal and social) reasons not to do so.

14. Our argument rests on the premise that organizations using AI for decision-making take steps to address the relevant issue of bias in training data.

15. https://www.uipath.com/product/rpa-ai-integration-with-ai-center (accessed October 3, 2023).

16. https://cvviz.com/product/resume-screening/ (accessed March 7, 2023).

17. https://www.nellyssecurity.com/blog/articles/video-surveillance/what-is-deep-learning-ai-and-why-is-it-important-for-video-surveillance (accessed November 2, 2023).

18. https://openai.com/blog/chatgpt (accessed January 23, 2023)

19. https://techhq.com/2019/04/ibm-could-be-a-model-for-hr-in-the-ai-age/ (accessed January 27, 2023).

20. https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html (accessed March 3, 2023).

21. https://api.slack.com/bot-users (accessed October 2, 2023).

22. https://www.accenture.com/gb-en/services/applied-intelligence/solutions-ai-pricing (accessed January 21, 2023).

References

Acar, O. A., & West, D. 2021. When an educated guess beats data analysis. *Harvard Business Review*, June. Retrieved from https://hbr.org/2021/06/when-an-educated-guess-beats-data-analysis.

Acemoglu, D., & Restrepo, P. 2019. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33: 3–30.

Adner, R., Puranam, P., & Zhu, F. 2019. What is different about digital strategy? From quantitative to qualitative change. *Strategy Science*, 4: 253–261.

Agrawal, A., Gans, J., & Goldfarb, A. 2018. *Prediction machines: The simple economics of artificial intelligence*. Boston, MA: Harvard Business Review Press.

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., & … Welling, M. 2020. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53: 18–28.

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

Argote, L. 2013. Organization learning: A theoretical framework. In L. Argote (Ed.), *Organizational learning: Creating, retaining and transferring knowledge*: 31–56. Boston, MA: Springer.

Armstrong, J. S. 2001. Combining forecasts. In J. S. Armstrong (Ed.), *International series in operations research & management science*: 417–439. Boston, MA: Springer US.

Asch, S. E. 1956. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9): 1–70.

Athey, S. 2018. The impact of machine learning on economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda*: 507–547. Chicago, Illinois: University of Chicago Press.

Attenberg, J., Ipeirotis, P., & Provost, F. 2015. Beat the machine. *Journal of Data and Information Quality*, 6(1): 1–17.

Baars, B., & Gage, N. M. 2013. *Fundamentals of cognitive neuroscience: A beginner's guide*. New York, New York: Academic Press.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. 2021. Beware explanations from AI in health care. *Science*, 373: 284–286.

Balasubramanian, N., Ye, Y., & Xu, M. 2022. Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47: 448–465.

Baldwin, C. Y., & Clark, K. B. 2000. *Design rules: The power of modularity (Vol. 1)*. MIT press.

Bandura, A. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84: 191–215.

Bao, X., Diabat, A., & Zheng, Z. 2020. An ambiguous manager's disruption decisions with insufficient data in recovery phase. *International Journal of Production Economics*, 221: 107465.

Battigalli, P., Francetich, A., Lanzani, G., & Marinacci, M. 2019. Learning and self-confirming long-run biases. *Journal of Economic Theory*, 183: 740–785.

Becker, J., Guilbeault, D., & Smith, N. 2021. *The crowd classification problem: Social dynamics of binary choice accuracy*. arXiv preprint:2104.11300.

Bertsimas, D., & Kallus, N. 2020. From predictive to prescriptive analytics. *Management Science*, 66: 1025–1044.

Bhardwaj, A., Yang, J., & Cudré-Mauroux, P. 2020. A human-AI loop approach for joint keyword discovery and expectation estimation in micropost event detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 2451–2458.

Bianchini, B., Halm, M., Matni, N., & Posa, M. 2021. *Generalization bounded implicit learning of nearly discontinuous functions*. arXiv preprint:2112.06881.

Bishop, C. M., & Nasrabadi, N. M. 2007. Pattern recognition and machine learning. *Journal of Electronic Imaging*, 16: 049901.

Blum, M., & Blum, L. 2021. A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, 08(01): 1–42.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., et al. 2022. *On the opportunities and risks of foundation models.* arXiv preprint:2108.07258.

Brown, G. 2010. Ensemble learning. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning*. Boston, MA, USA: Springer.

Brown, G., Wyatt, J., Harris, R., & Yao, X. 2005. Diversity creation methods: A survey and categorisation. *Information Fusion*, 6: 5–20.

Brynjolfsson, E., & Mitchell, T. 2017. What can machine learning do? Workforce implications. *Science*, 358: 1530–1534.

Burton, R. M., & Obel, B. 1984. *Designing efficient organizations: Modelling and experimentation*, Vol. 7. North Holland: Elsevier Science Publishers B.V.

Camillus, J. C. 2008. Strategy as a wicked problem. *Harvard Business Review*, 86: 98–101. Retrieved from https://hbr.org/2008/05/strategy-as-a-wicked-problem, Vol. 5.

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., & Smith, A. 2019. From soft classifiers to hard decisions. In D. Boyd, & J. Morgenstern (Eds.), *Proceedings of the conference on fairness, accountability, and transparency*: 309–318. New York, New York: ACM.

Cao, J., Udhayakumar, K., Rakkiyappan, R., Li, X., & Lu, J. 2021. A comprehensive review of continuous-/discontinuous-time fractional-order multidimensional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.

Carbonneau, R., Laframboise, K., & Vahidov, R. 2008. Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184: 1140–1154.

Christensen, M., & Knudsen, T. 2010. Design of decision-making organizations. *Management Science*, 56: 71–89.

Chui, M., Roberts, R., & Yee, L. 2022. *Generative AI is here: How tools like ChatGPT could change your business*. Quantum Black AI by McKinsey.

Cockburn, I., Henderson, R., & Stern, S. 2018. *The impact of artificial intelligence on innovation*. National Bureau of Economic Research. Working paper number 24449.

Condorcet, M. D. 1785. *Essay on the application of analysis to the probability of decisions rendered by a plurality of votes*. Paris: Imprimerie Royale.

Csaszar, F. A. 2012. Organizational structure as a determinant of performance: Evidence from mutual funds. *Strategic Management Journal*, 33: 611–632.

Csaszar, F. A., & Laureiro-Martínez, D. 2018. Individual and organizational antecedents of strategic foresight: A representational approach. *Strategy Science*, 3: 513–532.

Csaszar, F. A., & Ostler, J. 2020. A contingency theory of representational complexity in organizations. *Organization Science*, 31: 1198–1219.

Csaszar, F., & Steinberger, T. 2021. Organizations as artificial intelligences: The use of artificial intelligence analogies in organization theory. *Academy of Management Annals*, 16: 1–37.

Cyert, R. M., March, J. G., et al. 1963. *A behavioral theory of the firm*, Vol. 2: 169–187. Englewood Cliffs, New Jersey, US.

Dastin, J. 2018. *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved on 1st July 2022 from shorturl.at/dkvNU.

Daugherty, P. R., & Wilson, H. J. 2018. Humans plus robots: Why the two are better than either one alone. *Knowledge Wharton*, 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. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., … Hartig, F. 2018. Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88: 485–504.

Džeroski, S., & Ženko, B. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54: 255–273.

Floridi, L., & Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Garg, A., Shukla, N., Marla, L., & Somanchi, S. 2021. *Distribution shift in airline customer behavior during COVID-19*. arXiv preprint:2111.14938.

Gavetti, G., & Levinthal, D. 2000. Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly*, 45: 113–137.

Glikson, E., & Woolley, A. W. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14: 627–660.

Goodfellow, I., Bengio, Y., & Courville, A. 2016. *Deep learning*. Boston, Massachusetts, USA: MIT Press.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. 2013. *An empirical investigation of catastrophic forgetting in gradient-based neural networks*. arXiv preprint:1312.6211.

Graves, A., Wayne, G., & Danihelka, I. 2014. *Neural Turing machines*. arXiv preprint:1410.5401.

Grewatsch, S., Kennedy, S., & Bansal, P. 2021. Tackling wicked problems in strategic management with systems thinking. *Strategic Organization*, 21: 721–732.

Hastie, T., Friedman, J., & Tibshirani, R. 2009. *The elements of statistical learning*, Vol. 1. New York: Springer.

Hawkins, J. 2021. *A thousand brains: A new theory of intelligence*. New York, USA: Basic Books.

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

Heaven, D. 2019. Why deep-learning AIs are so easy to fool. *Nature*, 574: 163–166.

Heyes, C. 2018. *Cognitive gadgets: The cultural evolution of thinking*. Boston, Massachusetts, USA: Harvard University Press.

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

Hong, L., & Page, S. E. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101: 16385–16389.

Hu, W.-F., Lin, T.-S., & Lai, M.-C. 2022. A discontinuity capturing shallow neural network for elliptic interface problems. *Journal of Computational Physics*, 469: 111576.

Iansiti, M., & Lakhani, K. R. 2020. Competing in the age of AI. *Harvard Business Review*, 98: 60–67.

Ibrahim, R., Kim, S.-H., & Tong, J. 2021. Eliciting human judgment for prediction algorithms. *Management Science*, 67: 2314–2325.

Jain, S., Munukutla, S., & Held, D. 2019. Few-shot point cloud region annotation with human in the loop. *ICML Workshop*.

Janis, I. L. 1982. *Groupthink: Psychological studies of policy decisions and fiascoes*. Boston, Massachusetts, USA: Houghton Mifflin.

Jarrahi, M. H. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision-making. *Business Horizons*, 61: 577–586.

Kamar, E. 2016. Directions in hybrid intelligence: Complementing AI systems with human intelligence. In S. Kambhampati (Ed.), *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, 4070–4073. AAAI Press.

Kanet, J. J., & Adelsberger, H. H. 1987. Expert systems in production scheduling. *European Journal of Operational Research*, 29: 51–59.

Knudsen, T., & Srikanth, K. 2014. Coordinated exploration: Organizing joint search by multiple specialists to overcome mutual confusion and joint myopia. *Administrative Science Quarterly*, 59: 409–441.

Kratsios, A. 2021. The universal approximation property. *Annals of Mathematics and Artificial Intelligence*, 89: 435–469.

Krogh, A., & Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 7: 231–238.

Lazer, D., & Friedman, A. 2007. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52: 667–694.

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

Lee, W.-C., Chen, S., & Wu, C.-C. 2010. Branch-and-bound and simulated annealing algorithms for a two-agent scheduling problem. *Expert Systems with Applications*, 37: 6594–6601.

Le Roux, N., & Bengio, Y. 2010. Deep belief networks are compact universal approximators. *Neural Computation*, 22: 2192–2207.

Levitt, B., & March, J. G. 1988. Organizational learning. *Annual Review of Sociology*, 14: 319–338.

Lichtendahl, K. C., Grushka-Cockayne, Y., & Winkler, R. L. 2013. Is it better to average probabilities or quantiles? *Management Science*, 59: 1594–1611.

Lin, H., & Jegelka, S. 2018. *ResNet with one-neuron hidden layers is a universal approximator*. arXiv preprint:1806.10909.

Lindebaum, D., Vesa, M., & Hond, F. d. 2020. Insights from "the machine stops" to better understand rational assumptions in algorithmic decision-making and its implications for organizations. *Academy of Management Review*, 45: 247–263.

Lipton, Z. C. 2018. The mythos of model interpretability. *Queue*, 16: 31–57.

Llanas, B., Lantarón, S., & Sáinz, F. J. 2008. Constructive approximation of discontinuous functions by Neural Networks. *Neural Processing Letters*, 27: 209–226.

Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. 2022. News from generative artificial intelligence is believed less. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 86: 97–106.

Lönngren, J., & van Poeck, K. 2021. Wicked problems: A mapping review of the literature. *International Journal of Sustainable Development & World Ecology*, 28: 481–502.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science*, 2: 71–87.
McPherson, M., Smith-Lovin, L., & Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27: 415–444.
Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. 2012. Ensemble approaches for regression. *ACM Computing Surveys*, 45(1): 1–40.
Mintzberg, H. 1975. The manager's job: Folklore and fact. *Harvard Business Review*, July/August: 42.
Mintzberg, H. 1979. An emerging strategy of "direct" research. *Administrative Science Quarterly*, 24: 582.
Mollick, E. R., & Mollick, L. 2022. *New modes of learning enabled by AI chatbots: Three methods and assignments*. Available at SSRN.
Murray, A., Rhymer, J., & Sirmon, D. G. 2020. Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46: 0–44.
Myers, C. G. 2018. Coactive vicarious learning: Toward a relational theory of vicarious learning in organizations. *Academy of Management Review*, 43: 610–634.
Myers, C. G. 2021. Performance benefits of reciprocal vicarious learning in teams. *Academy of Management Journal*, 64: 926–947.
Nisbet, R., Elder, J., & Miner, G. 2009. Model complexity (and how ensembles help). *Handbook of Statistical Analysis and Data Mining Applications*: 707–721.
O'Hagan, A. 2006. *Uncertain judgements: Eliciting experts' probabilities. Statistics in Practice*. London; Hoboken, NJ: John Wiley & Sons.
Ostheimer, J., Chowdhury, S., & Iqbal, S. 2021. An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles. *Technology in Society*, 66: 101647.
Page, S. E. 2007. Making the difference: Applying a logic of diversity. *Academy of Management Perspectives*, 21: 6–20.
Page, S. E. 2010. *Diversity and complexity*, Vol. 2. New Jersey, USA: Princeton University Press.
Page, S. E. 2014. Where diversity comes from and why it matters? *European Journal of Social Psychology*, 44: 267–279.
Park, S., & Puranam, P. 2023. Vicarious learning without knowledge differentials. *Management Science*, Articles in Advance: 1–21.
Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. 2021. *Carbon emissions and large neural network training*. arXiv preprint:2104.10350.
Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., & Ben-Gal, I. 2020. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134: 113290.
Phillips, K. W., Northcraft, G. B., & Neale, M. A. 2006. Surface-Level diversity and decision-making in groups: When does deep-level similarity help? *Group Processes & Intergroup Relations*, 9: 467–482.
Piezunka, H., Aggarwal, V. A., & Posen, H. E. 2022. The aggregation–learning trade-off. *Organization Science*, 33: 1094–1115.
Polikar, R. 2012. Ensemble learning. In C. Zhang, & Y. Ma (Eds.), *Ensemble machine learning: Methods and applications*. New York, USA: Springer.
Puranam, P. 2021. Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design*, 10: 75–80.
Puranam, P., & Maciejovsky, B. 2020. Organizational structure and organizational learning. In L. Argote & J. M. Levine (Eds.), *The Oxford handbook of group and organizational learning*: 520–534. New York, USA: Oxford University Press.
Puranam, P., & Swamy, M. 2016. How initial representations shape coupled learning processes. *Organization Science*, 27: 323–335.
Raisch, S., & Krakowski, S. 2020. Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46: 192–210.
Ranjan, R., & Gneiting, T. 2010. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 71–91.
Reeve, H. W., & Brown, G. 2018. Diversity and degrees of freedom in regression ensembles. *Neurocomputing*, 298: 55–68.
Rittel, H. W. J., & Webber, M. M. 1973. Dilemmas in a general theory of planning. *Policy Sciences*, 4: 155–169.
Rougier, J. 2016. Ensemble averaging and mean squared error. *Journal of Climate*, 29: 8865–8870.
Sagi, O., & Rokach, L. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8: e1249.

Sah, R. K., & Stiglitz, J. E. 1985. The social cost of labor and project evaluation: A general approach. *Journal of Public Economics*, 28: 135–163.

Sah, R. K., & Stiglitz, J. E. 1986. The architecture of economic systems: Hierarchies and polyarchies. *The American Economic Review*, 76: 716–727.

Sah, R. K., & Stiglitz, J. E. 1988. Committees, hierarchies and polyarchies. *The Economic Journal*, 98: 451–470.

Samek, W., Wiegand, T., & Müller, K. R. 2017. *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. arXiv preprint:1708.08296.

Santos, F. M., & Eisenhardt, K. M. 2005. Organizational boundaries and theories of organization. *Organization Science*, 16: 491–508.

Scarselli, F., & Tsoi, A. C. 1998. Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, 11: 15–37.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning*, 5: 197–227.

Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. d., Vreede, G.-J. d., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. 2020. Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57: 103174.

Sen, P., & Puranam, P. 2022. Do alliance portfolios encourage or impede new business practice adoption? Theory and evidence from the private equity industry. *Strategic Management Journal*, 43: 2279–2312.

Settles, B. 2012. *Active learning*. Cham, Switzerland: Springer.

Shaw, P., Uszkoreit, J., & Vaswani, A. 2018. *Self-attention with relative position representations*. arXiv. preprint::1803.02155.

Shrestha, Y. R., Ben-Menahem, S. M., & Krogh, G. v. 2019. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61: 66–83.

Simon, H. A. 1996. The science of design: Creating the artificial. *The Sciences of the Artificial*. Cambridge, Massachusetts, USA: MIT Press.

Simons, T., Pelled, L. H., & Smith, K. A. 1999. Making use of difference: Diversity, debate, and decision comprehensiveness in top management teams. *Academy of Management Journal*, 42: 662–673.

Stephen, O., Sain, M., Maduh, U. J., & Jeong, D.-U. 2019. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, Article ID 4180949, 7 pages, 2019.

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119: e2111547119.

Surowiecki, J. 2005. *The wisdom of crowds*. New York, USA: Anchor.

Sutton, R. S., & Barto, A. G. 2018. *Reinforcement learning: An introduction*. London, England: A Bradford Book.

Thomas, E. A. C., & Ross, B. H. 1980. On appropriate procedures for combining probability distributions within the same family. *Journal of Mathematical Psychology*, 21: 136–152.

Torrey, L., & Shavlik, J. 2010. Transfer learning. In E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, & A. J. S. Lopez (Eds.), *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*: 242–264. Hershey, Pennsylvania, USA: IGI global.

Tschang, F. T., & Almirall, E. 2021. Artificial intelligence as augmenting automation: Implications for employment. *Academy of Management Perspectives*, 35: 642–659.

Tumer, K., & Ghosh, J. 1996. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8: 385–404.

Ueda, N., & Nakano, R. 1996. Generalization error of ensemble estimators. In M. Jorda, & T. Petsche (Eds.), *Proceedings of international conference on neural networks*, Vol. 1: 90–95. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. 2017. *Attention is all you need*: *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates.

von Hippel, E. 1990. Task partitioning: An innovation process variable. *Research Policy*, 19: 407–418.

Vellido, A. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32: 18069–18083.

Willis, J. 1996. A flexible framework for task-based learning. *Challenge and Change in Language Teaching*, 52: 62.

Zhang, C., & Ma, Y. 2012. *Ensemble machine learning: Methods and applications*. New York, USA: Springer.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., He, X., & He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109: 43–76.

Appendix 1: Deep Learning Applications for Managerial Tasks

Similar to any data-driven methodology, deep learning has its limitations. Training a deep learning model requires a large amount of data. In addition to limitations concerning data availability, the time required for training a model can also be extremely high, depending on its complexity. Furthermore, there could be constraints on computational power to fit a sufficiently complex model within reasonable time, and the desire to retain interpretability in human terms (Babic, Gerke, Evgeniou, & Cohen, 2021). Deep learning networks are also prone to adversarial attacks (e.g., with some pixel changes in an image, a lion can be recognized as a library with very high accuracy; Heaven, 2019) and affected by catastrophic

Table A1.1
Successful Practical Applications of Deep Learning to Managerial Tasks

Source	Managerial Task	Deep Learning Application
UiPath[15]	Expense (reimbursement) approval	Deep learning algorithms work alongside a human team to enter data (e.g., item, invoice number, amount) into the reimbursement system and then approve/reject it.
CVVIZ[16]	Resume screening	Algorithms parse and understand the contextual text in the resume and shortlist suitable candidates for a job opportunity.
Nelly's Security[17]	Prediction of unsafe movement	Nelly's Security employs deep learning tools to analyze movement of equipment and H to predict unsafe movement (e.g., H walking on a material-movement aisle in a warehouse). This technology is also implemented in warehouses to identify aggressive behavior of employees.
OpenAI[18]	Drafting of correspondence (e.g., letters, emails), summarizing documents, and generating reports	ChatGPT is an AI chatbot that OpenAI developed and launched in November 2022 that builds on a family of Large Language Models and fine-tuned with supervised and reinforcement learning.
IBM[19 20]	Employee career feedback, task allocation to employees based on skills	My Career Advisor (henceforth, MYCA) is an AI-based virtual assistant built on Watson (the question-answering computer system developed by IBM) to provide employees career feedback, especially about areas where they need to improve their skills. IBM uses MYCA in combination with Blue Match Technology, another AI-based solution that matches employees to tasks based on their AI-inferred skills.
Slack[21]	Information provision, coordination	Slack allows customers to embed into its channel bots that interact with users to provide them with information and coordination.
Accenture[22]	Pricing	Solution.AI is an AI-based strategic pricing tool that ensures price optimization relying on real-time insights derived from market signals, competitive intelligence, and dynamic customers' preferences and willingness to buy. The tool can be used for personalized and dynamic pricing, revenue growth management, and to improve customer profitability

forgetting (Goodfellow, Mirza, Xiao, Courville, & Bengio, 2013). Function discontinuity further hampers learning as approximated functions show high error rates around the discontinuities, particularly for algorithms relying on backpropagation of errors as in deep learning. We refer the interested reader to Table A3.1 in Appendix 3, which provides a glossary of key technical terms used in the paper, including those related to the limitations of deep learning.

Despite these limitations, deep learning has been effectively used in practice and technical advances are being made to tackle these challenges. For instance, to reduce the time and data required for model training, researchers are developing compact models and transfer learning methods (Torrey & Shavlik, 2010). Similarly, to address network forgetting, Neural Turing Machines (Graves, Wayne, & Danihelka, 2014) are being developed that could possibly resemble human-like memory. Novel adversarial training methods also hold the promise to overcome the issue of adversarial attack issues. Finally, in practice, continuous functions often turn out to be a satisfactory substitute for discontinuous ones, rendering neural networks applications broadly feasible. For instance, the robotics community has developed novel architectures (Bianchini, Halm, Matni, & Posa, 2021) that display favorable performance under discontinuity, and discontinuity-capturing neural networks (Hu, Lin, & Lai, 2022) also seem to perform well in this context. Other workarounds are also available in practice (e.g., piecewise continuous functions, functional approximations; see Llanas, Lantarón & Sáinz, 2008 for an example). Finally, more complex networks (i.e., three layers neural networks) hold the promise of being able to represent even discontinuous functions (Llanas et al., 2008). Table A1.1 illustrates successful applications of deep learning to managerial tasks.

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 in 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 organizations. 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 Negative Correlation, or 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).

Appendix 3: Glossary of Key Technical Terms Used in the Paper

Table A3.1
Glossary of Key Technical Terms

Term	Definition
Adversarial Attack	Purposefully generate challenging (adversarial) examples to improve model's prediction reliability.
Augmentation	AI is added to a human agent in decision-making.
Automation	AI replaces human agents in decision-making.
Backpropagation	Backpropagation is an algorithm which helps in supervised learning by calculating the gradient of the error function with respect to neural network's weights, and adjusts the weights iteratively to minimize the error.
Bagging	Random samples of the training data are selected with replacement. Using the random samples, independent models are trained in parallel, and their outcomes are aggregated.
Boosting	Models are trained sequentially, using higher weights for the mistakes of the previous model.
Catastrophic Forgetting	Catastrophic forgetting is a tendency where, while training a neural network, it could drastically and abruptly forget previously learned information upon learning new information.
Deep Learning	Learning with a neural network with three or more layers.
H-AI Ensemble	An ensemble (division of labor without specialization) between H and AI.
Hyperparameter	A parameter whose value is used to control the learning process and derive the model parameters.
NC Learning	Negative correlation learning refers to an ensemble learning technique that attempts to train and combine individual neural networks into the same learning process with the goal of generating the best result for the entire ensemble.
Neural Network	Neural networks are subset of ML that mimic the human brain and are comprised of an input layer, one or more hidden layers, and an output layer.
Stacking	Different types of models are trained using the same training data, for example, decision tree and linear regression. Finally, their outcomes are aggregated.
Task	A goal-oriented activity to achieve a desired outcome. Different tasks have different outputs that satisfy different goals.
Tuning	Finding the optimal values of the hyper parameters.
Weak Learner	An agent/model that makes predictions that are better than random guesses.