Journal of Organization Design (2021) 10:75–80
https://doi.org/10.1007/s41469-021-00095-2

POINT OF VIEW

Human–AI collaborative decision-making as an organization design problem

Phanish Puranam[1]

Received: 11 May 2020 / Accepted: 4 February 2021 / Published online: 16 February 2021

Abstract
The promise of collaboration between humans and algorithms in producing good decisions is stimulating much experimentation. Drawing on research in organization design can help us to approach this experimentation systematically. I propose typologies for considering different forms of division of labor between human and algorithm as well as the learning configurations they are arranged in, as basic building blocks for this endeavor.

Keywords AI · Organization design · Machine learning · Learning configurations · Division of labor

Introduction

In this brief *Point of View* article, I offer some thoughts on how we may conceptualize collaborative decision-making between humans and AI algorithms as a problem in organization design.

While there are many possible forms of interaction between humans and AI algorithms, the arguments here are most relevant to knowledge work in which humans and AI algorithms through some form of collaboration, together produce a decision that is implemented by a third party (for instance stock picking, investment, sentencing, screening candidates). I refer to these as situations of "Human–AI Collaborative Decision-Making" (or HACD). The arguments may also apply to situations which involve a human training an AI algorithm (e.g., self-driving cars that learn from observing humans drive) or vice versa (e.g., chatbot-based language learning applications), or the use of algorithms to improve matches between humans (e.g., friend suggestions on social media platforms), but will require additional considerations that I do not address here.

In what follows, I use the terms "AI" and "algorithm" interchangeably with "machine learning" (ML). I am aware that not all algorithms are AI, and not all AI is machine learning (Broussard 2018; Raj and Seamans 2019), but my usage avoids tedium. Human–ML collaboration for decision-making is different from other forms of human interaction with/adoption of technology for at least two reasons. First, the outputs of decision tasks unlike physical production tasks can sometimes be aggregated in a manner that improves their accuracy through error cancellation. This means that division of labor without specialization is feasible in HACD. Second, there is the potential for mutual adjustment: both humans and AI based on machine learning algorithms are adaptive systems that change how they make decisions over time through learning from experience (i.e., past data). Organizational scientists understand the dynamics of mutual adaptation and know them to be distinctive from one sided-adaptation or static collaboration (Knudsen and Srikanth 2014; Lave and March 1993; Lounamaa and March 1987; Puranam and Swamy 2016).

I make two main points in this article. First, I note that from the perspective of organization design, there are multiple possible configurations of division of labor in HACD besides the one that is most intuitive and currently dominates popular discourse, namely one based on specialization (human and AI each do different sub-tasks they are relatively best at). Second, organization design research also suggests that there are multiple learning configurations to consider in which humans and AI may "learn together". I express these possibilities in terms of simple typologies. Together they describe a design space which may not yet be completely or densely populated, but which may serve to guide our explorations in terms of both practice and theory.

✉ Phanish Puranam
Phanish.puranam@insead.edu

[1] INSEAD, 1 Ayer Rajah Avenue, Singapore, Singapore

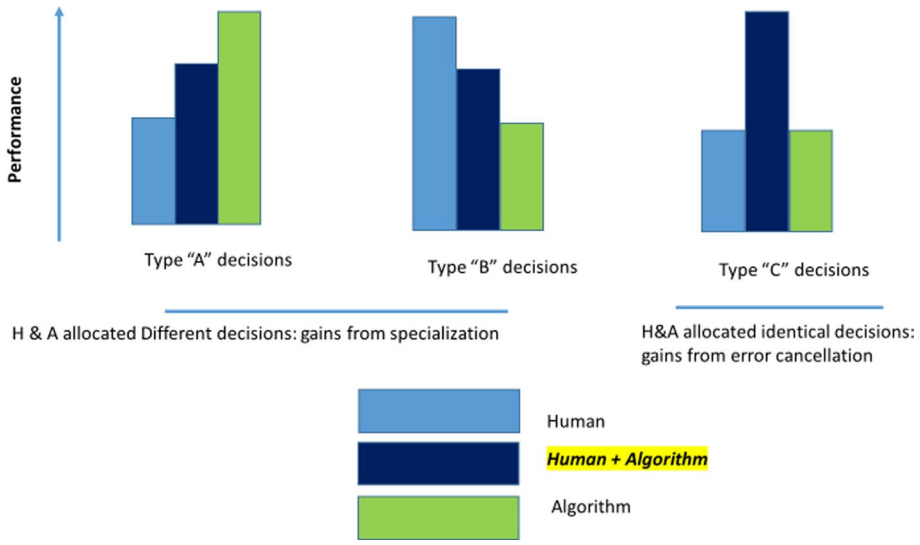

Fig. 1 Gains from human–algorithm collaborative decision-making (HACD)

How HACD can be valuable

Taking an organization design perspective on the problem of human–AI collaborative decision-making (henceforth HACD) requires us to view the combination of the human and the algorithm as an organization—i.e., a multi-agent, goal-oriented system. The goal of an HACD organization is to produce a decision. The design of the organization constitutes the choices about division of labor (task division and task allocation) and integration of effort (information and reward provision, exception handling) that characterize the organization (Puranam 2018).

Why should an organization that involves HACD ever be superior to an organization comprising only humans or only algorithms? A first cut at the problem is shown in Fig. 1.

At any point in time, given the prevailing state of technology, we can speak of three types of tasks: Type A tasks are those in which algorithms equal or outperform humans (e.g., today that would include image or handwriting recognition). Type B tasks are the ones that humans outperform algorithms on (e.g., evaluating a job applicant's integrity remains a Type B task, even though reading the CV might be a Type A task) and must remain in the hands of humans. However, by appropriate division of labor—including breaking up aggregate tasks into smaller ones which can be differentiated into Type A and Type B[1]—the human and algorithm each can do what they are better at. This unlocks gains from specialization (net cost of coordination between agents) through HACD. This logic lies at the heart of a lot of contemporary workflow and process automation. It is not fundamentally different from the calculus of outsourcing or offshoring or gains from trade as set out by Adam Smith and David Ricardo; and it applies to tasks in general, not only decisions.

The more interesting case is Type C, where despite no clear superiority of either human or algorithm, the combination through aggregation may outperform either alone. A distinctive feature of decisions is that their accuracy can sometimes be improved through pooling and error cancellation (Larrick and Soll 2006; Rokach 2010; Surowiecki 2004). Having a human and an algorithm (or indeed several algorithms—as is the case with ensemble learning models) make the same decision and then aggregating their outputs can produce improved quality in terms of greater decision accuracy. This is not possible usually with physical products. For decisions that involve predicting a continuous variable (e.g., quality), the "wisdom of crowds" provides the intuition. For decisions that involve predicting a discrete category (e.g., hire or reject), Condorcet's jury theorem provides a foundation, which illustrates how increasing the size of a jury of identically and modestly accurate members can increase the jury's aggregate accuracy.

Possible divisions of labor between humans and AI in collaborative decision-making

Building on this intuition of the differences between types of tasks, and by drawing on basic ideas in organization design, we can give a more comprehensive picture of the possible divisions of labor in HACD. Figure 2 illustrates the arguments below by giving hypothetical divisions of labor between humans and AI in the context of HACD for stock picking, such that an equity analyst and an algorithm might jointly make a recommendation on whether to buy a stock.

[1] I see the situations where algorithms handle routine cases (Type A) and humans the exceptions (Type B) also as an instance where a task can be partitioned into sub-tasks that fall into Type A or Type B categories.

	Parallel	Sequential
Specialize to make different types of decisions (i.e. different inputs and outputs)	H does analysis of qual data, A does analysis of quant data; final stock recommendation report contains both components	A does processing of quant data, H integrates that with insight from qual data or reverse order; final stock recommendation comes from A (or H)
Make the same type of decision	H and A make price forecasts independently on same data, stock recommended if either agree (or alternately only if both agree)	H and A make price forecasts one after the other on same data, but the second only sees if the first approves; stock recommended only if H or A finally agrees

Fig. 2 Possible divisions of labor for stock picking with HACD

A division of labor involves (a) decomposition of the goal (the final decision) into tasks (decisions) that aggregate into the final decision and (b) allocating sub-clusters of these tasks across the organization's members. The resulting allocation of tasks to agents—*a* division of labor—can be described in two ways.

First, there will exist a structure of interdependence between tasks (and therefore between clusters of tasks allocated to different agents). Two tasks are interdependent if the value created when both tasks are performed is different from the sum of values created by performing each task alone. For instance, this could be because they draw on common rivalrous inputs, the value of outputs is super or sub-additive, or one is an input to another (Burton and Obel 1984; Milgrom and Roberts 1990; Thompson 1967). Since decision tasks do not usually consume tangible inputs, the relevant forms of interdependence between two decisions are the cases where one is a *sequential* input to the other (perhaps repeatedly, as in reciprocal interdependence),[2] or the value of their joint, *parallelly* produced outputs is super- (or sub) additive (also see Christensen and Knudsen 2013).

Second, the allocated tasks to agents may vary in the extent of heterogeneity of knowledge or skills needed in these tasks (Raveendran et al. 2020). Two workers who must produce dining tables may both produce a table each—a case of non-specialized task allocation—or focus differently and, respectively, on making legs and tops (an object-based division of labor) or in cutting and fixing wood (an activity-based division of labor). It is by no means obvious which of these arrangements is superior, as it depends on the gains from *specialization* (by each worker focusing on a narrow set of tasks they are distinctively good at) vs. the gains from *customization* (i.e., managing the dependencies between dissimilar task), as well as the cost of coordination among agents (Raveendran et al. 2015). The difference between craft and industrial production of furniture illustrates this point.

The gains from specialization in parts of a decision (splitting into Types A and B), whether in sequence or in parallel, thus constitute but one form of HACD. The gains from ensembling, i.e., allowing multiple agents to make the identical decision, may apply to Type C tasks. Of course, tasks might change from one Type to another over time as technology advances, perhaps inevitably in the direction of Type A by depleting Types B and C—but it is enough for my arguments that each Type exists at any point in time. Persistent data limitations and the possible non-stationarity of the underlying data generation processes can prevent algorithms (or humans) from accomplishing outright superiority, possibly making Type C a stable category. For instance strategic decisions may have these attributes.

In sum, the division of labor in HACD can be described along two dimensions: the nature of interdependence—whether the decisions of the human and algorithm are related sequentially (only one of their outputs matters directly for final output) or can occur in parallel (human and algorithm outputs both directly matter for final output), and the nature of specialization—whether they engage in different or identical decision tasks.

From static to dynamic considerations: learning configurations within HACD

So far, we have considered a rather static picture of the division of labor between humans and AI for decision-making—which simply assumes differences in what they are good at. As Adam Smith pointed out division of labor not only exploits existing differences in skill in allocating different components of labor to different actors, but the different

[2] Note that the notion of decision rights (which agent can choose to accept/reject output of other) can be treated the same as who is last in a sequence.

Table 1 Learning configurations in human–AI collaborative decision-making (HACD)

	Communication constraints	Communication is feasible *on inputs/process/outputs/ feedback*
Independent feedback	Isolated learning	Vicarious learning
Interdependent feedback	Coupled learning	Coupled + vicarious learning

allocations themselves produce difference in skill over time (also see Mintzberg 1983 for an elaboration of this point in the context of organization design). Further, the distinctive feature of HACD, as opposed to other forms of technology adoption or even automation, is the potential for mutual adjustment: both humans and algorithms not only learn on their tasks from feedback, they also learn to adjust *to* each other and *from* each other.

Learning here refers to a change in beliefs or behavior (not necessarily performance improvement) as a consequence of experience (Argote 2013). Learning in the context of decision-making implies that given the same input at two different points in time, a decision-maker (either human or algorithm) may produce different outputs (i.e., take different decisions), because of changes to how the inputs are processed that occurred in the intervening period. These changes are the result of feedback conditional on the output, which is itself conditional on inputs. For an isolated human decision-maker, the data needed to learn how to make decisions should therefore necessarily include *feedback/evaluation* of past decisions conditional on the *output* (the actual decision they made) as well as the *inputs* they based their decision on, and possibly the *process* they used to arrive at a decision (the last may not be necessary given sub-conscious decision-making and associative learning).

Therefore, to understand how members of a HACD organization learn, it is useful to ask what might be different about the data available to them in terms of feedback conditional on inputs, outputs and process, compared to the case where they acted as isolated decision-makers. I use the term *learning configurations* to characterize situations that vary in terms of the nature of information available for learning. The organization design literature suggests two dimensions (Table 1) on which learning configurations might vary in situations of multi-agent learning.

The first is interdependence between the decision-makers. Organization designers recognize the important distinction between interdependence between tasks (in this case, decisions which we have described in terms of parallel or sequential) vs interdependence between agents (Puranam et al. 2012). Given two tasks undertaken by agents A and B, (symmetric) interdependence between agents exists when the value of A's actions to A depend on B's actions and vice versa (Emerson 1962; Kelley and Thibaut 1978; Pfeffer and Salancik 1978; von Neumann and Morgenstern 2007). We can observe interdependence between agents even when there is none between the task they perform or vice versa. In HACD, if the feedback to A on A's decisions depends on B's decisions and vice versa, then they are interdependent—and their learning will be *coupled* (Lave and March 1993; Lounamaa and March 1987; Knudsen and Srikanth 2014; Puranam and Swamy 2016).

For instance, in a HACD organization of one human and one algorithm that together produce an equity research report, we might provide feedback separately on the components of the report that the human and the algorithm contributed or on the report as a whole (was it good or bad). In the second situation, the human and algorithm are coupled in their learning, because the feedback they receive is on the aggregate output but not in the first (though the decisions they make are interdependent in both cases). This is akin to the distinction between carpenters who receive feedback on the whole table they produce ("how much did the customer pay?") or on the parts they contributed ("beautiful finish on the surface! rickety legs though").

Second, situations vary in the ease with which agents can *communicate*—exchange information on the inputs and process they use to decide, as well as the decision themselves. This is not necessarily a matter of all or nothing. Communication is particularly difficult even among human decision-makers who specialize in different tasks (Dougherty 2001). Between humans and algorithms, it may be difficult because the sheer volume of information overloads human capacities—for instance, when the algorithm is used as a screening device, making it difficult for the human to process even the inputs and outputs that algorithm produced. It may also be hard to exchange information on the processes used to decide as highlighted by the literature on the challenges of building explainability into AI (Samek et al. 2017). However, to keep the exposition simple, I consider all cases where some communication between human and algorithm is possible as instances of *vicarious* learning: one agent learns from the experience of another, where experience may be any combination of past inputs, process, outputs and feedback (Bandura 1977; Cyert and March 1963). For instance in a HACD team of one human and one algorithm that produce a recommendation on an equity, the human may have access to the inputs (data) and outputs (results) produced by the algorithm and vice versa, or not; the latter represents a situation of communication constraints (these need not be symmetric

of course). In our example of physical good production, the carpenters might see the feedback each receives as well as the inputs and raw materials each uses (or not).

Learning configurations in HACD can therefore range from isolated learning (independent feedback, no communication with other members) to situations involving both coupled and vicarious learning (interdependent feedback, with communication between members), or either alone. However, it is important to highlight that in all cases, the decisions themselves could be interdependent. Further, in all cases (including isolated learning), mutual adjustment between human and algorithm could be taking place, if the division of labor between humans and algorithms affects what data are available to each. For instance, two bank officers who decide on mortgage applications and learn from individual feedback on past cases with no communication between them, may still be tacitly adjusting to each other when placed in a serial division of labor, because the learning opportunities of the downstream agent depend on the actions of the upstream agent (Christensen and Knudsen 2020).

Combining division of labor and learning configurations: the design space for HACD

Considering HACD organizations with a joint emphasis on the nature of division of labor and the learning configuration can help us understand and design them better, both in terms of expanding the space of possibilities, as well as the precision with which we characterize particular points within them.

Some models already exist in the organization design literature for each of the types of possible division of labor in HACD in at least some of the possible learning configurations. These models typically use adaptive reinforcement learning algorithms to simulate human decision-makers—but that should not prevent us from re-interpreting them as models of HACD, particularly once heterogeneity between agents is added to the picture.

For instance, when the division of labor in decision-making involves specialization, feedback is often on group level output. Models of coupled learning have highlighted that the key design challenge in such situations is to avoid superstitious learning from false negatives and false positives (Lave and March 1993; Lounamaa and March 1987). Common priors and vicarious learning in the case of parallel specialized decisions (Puranam and Swamy 2016; Knudsen and Srikanth 2014; Aggarwal et al. 2017), and the stability of personnel in the case of sequential specialized decisions (Denrell et al. 2004) have been argued to mitigate the challenge. Coupled learning might also arise without specialization. Piezunka et al. (2020) study learning by participation, in which parallel unspecialized decision-makers receive feedback only on their aggregate decisions derived from voting. They point out that the quality of decisions over time depend on how the contrarians—those whose beliefs did not align with the majority vote at a point in time—influence future decisions. We also know that serial and parallel architectures lead to different learning dynamics even with isolated learners, because the inputs and therefore opportunities for learning are censored in serial architectures (Christensen and Knudsen 2013, 2020).

However, these hardly exhaust the combinatorial space obtained by crossing possible divisions of labor with learning configurations. There is much to do in terms of completing our conceptual understanding of these possibilities, and even more to do in confronting the models with data. A partnership between organization design researchers and practitioners interested in HACD seems ripe with opportunity.

Acknowledgements I thank Bart Vanneste, Marlo Raveendran, Sanghyun Park, Yash Raj Shreshtha and Thorbjorn Knudsen for helpful suggestions.

Authors' contributions Phanish Puranam is the sole author. The author read and approved the final manuscript.

Funding Funding from INSEAD R&D Committee is gratefully acknowledged for research assistance.

Availability of data and materials Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests The author declares that there are no competing interests.

References

Aggarwal VA, Posen HE, Workiewicz M (2017) Adaptive capacity to technological change: a microfoundational approach. Strateg Manag J 38(6):1212–1231. https://doi.org/10.1002/smj.2584

Argote L (2013) Organizational learning: creating, retaining and transferring knowledge. Springer US, New York

Bandura A (1977) Social learning theory. Prentice Hall, Englewood cliffs

Broussard M (2018) Artificial unintelligence: how computers misunderstand the world. MIT Press, Cambridge
Burton RM, Obel B (1984) Designing efficient organizations: modeling and experimentation.
Christensen M, Knudsen T (2013) How decisions can be organized-and why it matters. J Organ Des 2(3):41–50
Christensen M, Knudsen T (2020) Division of roles and endogenous specialization. In: Industrial and Corporate Change, pp 105–124
Cyert RM, March JG (1963) A behavioural theory of the firm. Wiley, Hoboken
Denrell J, Fang C, Levinthal DA (2004) From T-mazes to labyrinths: learning from model-based feedback. Manage Sci 50(10):1366–1378. https://doi.org/10.1287/mnsc.1040.0271
Dougherty D (2001) Reimagining the differentiation and integration of work for sustained product innovation. Organ Sci 12(5):612–631. https://doi.org/10.1287/orsc.12.5.612.10096
Emerson RM (1962) Power-dependence relations. Am Sociol Rev 27(1):31–41
Kelley HH, Thibaut JW (1978) Interpersonal relations: a theory of interdependence. Wiley, Hoboken
Knudsen T, Srikanth K (2014) Coordinated exploration: organizing joint search by multiple specialists to overcome mutual confusion and joint myopia. Adm Sci Q 59(3):409–441. https://doi.org/10.1177/0001839214538021
Larrick RP, Soll JB (2006) Intuitions about combining opinions: misappreciation of the averaging principle. Manage Sci 52(1):111–127
Lave CA, March JG (1993) An introduction to models in the social sciences. The University Press of America, Lanham
Lounamaa PH, March JG (1987) Adaptive coordination of a learning team. Manage Sci 33(1):107–123
Milgrom P, Roberts J (1990) The economics of modern manufacturing: technology, strategy, and organization. Am Econ Rev 80(3):511–528
Mintzberg H (1983) Structure in fives: designing effective organizations. Prentice Hall PTR, Upper Saddle River
Pfeffer J, Salancik GR (1978) The external control of organizations: a resource dependence perspective. Stanford University Press, Palo Alto
Piezunka H, Aggarwal VA, Posen HE (2020) Learning-by-participating: the dual role of structure in aggregating information and shaping learning. Organ Sci (forthcoming)
Puranam P (2018) The microstructure of organizations. Oxford University Press, Oxford
Puranam P, Swamy M (2016) How initial representations shape coupled learning processes. Organ Sci 27(2):323–335. https://doi.org/10.1287/orsc.2015.1033
Puranam P, Raveendran M, Knudsen T (2012) Organization design: the epistemic interdependence perspective. Acad Manag Rev 37(3):419–440. https://doi.org/10.5465/amr.2010.0535
Raj M, Seamans R (2019) Primer on artificial intelligence and robotics. J Organ Des 8(11):1–14
Raveendran M, Puranam P, Warglien M (2015) Object salience in division of labor: experimental evidence. Manage Sci 9:337–392
Raveendran M, Silvestri L, Gulati R (2020) The role of interdependence in the micro-foundations of organization design: task, goal, and knowledge interdependence. Acad Manag Ann 14(2):828–868
Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1–2):1–39. https://doi.org/10.1007/s10462-009-9124-7
Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (eds) (2017) Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature, New York
Surowiecki J (2004) The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. Doubleday & Co, New York
Thompson JD (1967) Organizations in action: social science bases of administrative theory. Transaction Publisher, Piscataway
von Neumann J, Morgenstern O (2007) Theory of games and economic behavior. Princeton University Press, Princeton