Healthcare and Medical Devices, Vol. 171, 2025, 277–286
https://doi.org/10.54941/ahfe1006214

A Tool to Complement Human Intelligence: The Math Behind Human Indispensability

Chris Robinson and Joshua Lancaster

Independent Researcher, Louisville, KY 40258, USA

ABSTRACT

Much ink has been spilled recently on the existential risks and potential of Artificial Intelligence. Between breathy utopian think-pieces and apocalyptic proclamations of the end of meaning in human life, an entire spectrum of outlooks muddies the waters on insight-driven and human-focused paths forward. While philosophical musings and abstract plans are prevalent, relatively little attention has been paid to underwriting integrative deployment as a problem which yields to analysis. The question 'when should an autonomous system step in' is typically framed as demanding a comprehensive world-model of the human subject- oppositional defiance and counter-picking make this approach undesirable, turning the human and AI against one another. Instead, by combining operationalization from psychology, Pareto optimality from economics, norm-based stability from robust controls, and shortest-path algorithms from graph theory, we are able to present mathematically robust conditions under which heterogenous systems provide superior performance to unitary agents, guaranteeing a lower bound on efficacy of joint human/AI teams endorsed by relative advantage. We also derive implicit conditions under which such relationships hold, finding them to be of geometrically increasing scope as task complexity increases. Finally, we demonstrate these relations are not merely theoretical, using sample tasks with adversarial complexity to challenge the assignment paradigm, and find the results to remain within an order-of-magnitude of the predicted robustness condition.

Keywords: Human/AI interaction, Human-centered robotics, Design and human factors

INTRODUCTION

Research on integration between humans and autonomous systems has consistently shown that joint teams exhibit superior performance to homogeneous ones (Crandall, 2002). Most interesting through our lens, these reports exist across widely varying domains: (Feng, 2016) for UAV control; (Sharma, 2023) in mental health services; (Hitsuwari, 2023) in poetry. (Li, 2024) outlines how the effectiveness of heterogeneous teams is one of the key factors enhancing their acceptability to users.

However, there are notable hurdles to adoption of such approaches, in spite of the demonstrated benefits. (Vassilakopoulou, 2023) implies inertial thinking is one such limit. (Ulfert, 2024) suggests novel integration mechanisms are likely necessary. Supplying rigorous and well-founded

rationale which endorse the advantages of Human/AI teams is therefor critical for guidance of AI adoption towards the benefit of humanity as a whole.

Several investigations of Human/AI interactions in more abstract contexts suggest causativ mechanisms: brainstorming (Memmert, 2023), military decision-making (Vold, 2024), and social collaboration (Westby, 2023), highlight critically that AI systems assist in reducing cognitive load on humans, precisely as other information technologies have historically.

Though there is ample work on engendering this effect, less examines the robustness of it. (Aghion, 2017) addresses this concept in an economic sense. Notably, they leverage the concept of relative advantage to suggest that asymmetric displacement of workers by automation systems can be counter-intuitively harmful to industries- a pattern observed with other forms of automation, already.

There is generally an optimum which can be achieved by examining the combinations of autonomous and human controls (Nichols, 2015). Comparable to the economic concept of 'comparative advantage', when robots and humans have asymmetric competencies, even if one agent has an absolute performance advantage over the other performance gains are possible when dividing total labor (Kim, 2011). A rigorous discussion of identifying autonomy levels can be found in (Barber, 1999). Notably, (Fu, 2015) presents a method using temporal logic to construct Pareto-Optimal policies, indicating that such optimal policies must exist in mixed autonomy systems.

Based on the results in these works we are able to infer, in conjunction with the reported efficacy of joint Human/AI systems, that there is an optimization problem embedded in the task of analyzing a mixed autonomy system. Knowing that this is the case, we can proceed with our analysis of the joint team as a problem of identifying conditions under which a non-extrema of autonomy level (all machine, or all human) produce optimal solutions.

In this paper, we use similar reasoning, framed in terms of the analysis of planning and levels of autonomy in a joint system, to address the deeper question of the robustness of the observed effect whereby heterogeneous systems outperform homogeneous ones. By considering the joint task as a graph of subtasks, we are able to apply concepts from robust controls to derive conditions under which this effect holds, and incidentally indicate the breadth of them. This, critically, demonstrates that the observed performance advantages of the heterogeneous are not outliers, but in fact the norm.

VARIABLE AUTONOMY

We will consider an arbitrary task, in the abstract, to be modeled by a task graph comprised of subtasks, noting that any sub-task may theoretically be decomposed in the same way. Our objective, then, is to optimize the expected cost for achieving the objective with respect to the path from the starting state to the goal state. As such, we are able to evaluate a locally optimal path rather than every possible solution, and determine conditions when this locally optimal path is also globally optimal.

Definitions

We presume a task which proceeds from some start state, S, to a goal state G with an intervening set of **subtasks** drawn from K-many in total, which effect the transition from S to G. The relationships between the different subtasks is represented by the task graph, T_G, a directed graph in which each vertex represents a subtask and the edges represent transitions from one subtask to another. We assume that this graph may contain many variable paths between S and G, and be hierarchically ordered.

Each subtasks may be accomplished by a human via at least one **manual module** M_i, or by the autonomous system via at least one **autonomous module** A_i. As each subtask in T_G has at least two modules that may be used, the number of states is doubled, and the number of paths to be evaluated grows as $O(2^K)$. As such, the issues associated with the computation of MDP policies can be clearly seen.

Our goal is determine which subtasks (denoted abstractly as P_i, note that the i index here does not refer to execution order, but is instead a subtask label) should be executed by an operator and which should be executed autonomously. The **autonomy level**, denoted A_L for this algorithm, parametrizes how many subtasks are executed autonomously. In this model, achieving the task means progressing the system from the initial state S via a subset of subtasks from the start to the goal. In this way, we can represent a solution for the task as a chain in T_G.

We seek the optimal solution *given* some A_L. As there are K-many subtasks, we will have $K + 1$ levels in total, including $A_L = 0$ at which all subtasks are marked for manual execution.

We also define an **autonomy level plan**, ρ to be a selection, for every subtask in T_G, whether the subtask will be executed autonomously or manually. Each autonomy level plan represents which subtasks will be executed autonomously and which manually. For example, the autonomy level plan for $A_L = 2$ for the T_G represented in Figure 1 might be $\rho_2 = \{A,M,M,M,A,M\}$

Note that a *plan* and a *solution* are not synonymous. The problem of determining the optimal autonomy level plans for a task is posed as the optimal order in which modules are toggled from manual execution to autonomous execution- recognizing that the process of identifying an optimal solution with these allocations is a distinct combinatoric problem.

Cost Function

In this section, we define a **cost**, μ, such as completion time, probability of success, operator fatigue, etc. which determines relative efficacy of task performance. $\mu(M_i)$ represents the cost associated with the manual module associated with the i^{th} subtask, $\mu(A_i)$ the same for the autonomous.

Let X, Y, or Z be arbitrary modules of either manual or autonomous type. Additionally, $\cdot$ is the operator representing the method of combination of multiple μ costs into a joint cost. We stipulate that μ meets the following:

Condition 1: Costs are non-negative, as otherwise it may be possible to construct an autonomy level plan with infinitely decreasing cost; $\mu(\mathrm{X}) \geq 0$.

Condition 2: When combining costs, the joint costs are monotonically increasing or decreasing: $\mu(\mathrm{X}) \cdot \mu(\mathrm{Y}) \geq \max(\mu(\mathrm{X}), \mu(\mathrm{Y}))$ **or** $\mu(\mathrm{X}) \cdot \mu(\mathrm{Y}) \leq \min(\mu(\mathrm{X}), \mu(\mathrm{Y}))$.

Condition 3: The cost function is transitive, ensuring that the effect of a module's independent cost on the cost of any autonomy level plan is the same; $\mu(\mathrm{X}) \cdot \mu(\mathrm{Y}) = \mu(\mathrm{X}) \cdot \mu(\mathrm{Z}) \rightarrow \mu(\mathrm{Y}) = \mu(\mathrm{Z})$.

Joint Expected Cost

A direct approach would be to evaluate the least-cost path for every possible ρ. However, this would require evaluating 2^{K} variants, which is computationally intractable. To ameliorate this, we evaluate whether each module should be executed manually or autonomously on basis of the effect doing so has on the cost as seen from the starting state, using an expected cost model.

In this model, grouped transitions between subtasks represent outcomes that are related in probability. Figure 2 shows the set of transitions leading from the example subtask N to either M or P, with probability of the event and expected cost of the result, and a mechanism for incorporating subtask failures, a **failure cost** $\mathrm{C_{FNX..Z}}$. Using this model, we can construct an alternative expected costs of subtask execution, denoted as *E(N|N..Z)*, including not only $\mu(N)$, but also a combination of C_{Nx}, which are the expected costs for the subsequent subtasks:

$$(E(N) \mid NX..Z) = \mu(N) + \sum_{\forall x \in X..Z} p_{Nx} E(x) + \left(1 - \sum_{\forall y \in X..Z} p_{Ny}\right) C_{F_{NX..Z}} \quad (1)$$

To then calculate the joint cost associated with the subtask N itself, we can utilize the minimum such calculated expectation:

$$E(N) = min\left(\{E\left(N \mid NX_1..Z_1\right), E\left(N \mid NX_2..Z_2\right), \ldots\}\right) \quad (2)$$

We can resolve this calculation by working backwards in T_G from the goal. G itself has no subsequent subtasks and is the terminal state, so its expected cost can be considered zero. If $E(S_{AL})$ represents the expected cost at S, given an autonomy level A_L, the cost change to a higher autonomy level $\mathrm{A_{L+1}}$ can be written as:

$$\Delta E(S)_{A_L} = E\left(S_{A_L+1}\right) - E\left(S_{A_L}\right) \quad (3)$$

we can then identify which subtask P_i will have the *most negative* change in cost when automated, and update the next autonomous level plan accordingly:

$$\rho_{A_{L+1_i}}[i] = \min_{A_l} \left\{\Delta E(S)_{A_L}\right\} \quad (4)$$

The change in cost $\min\{\Delta \mathrm{E(S)_{AL}}\}$ should be negative, as we seek to find an optimal autonomy plan, however, it is possible that some point, all remaining changes induce an increase in expected cost. This point is naturally the local minimum in ΔE(S), and we define it to be the optimum autonomy level.

ANALYSIS

In the prior section, we utilized the joint expected cost at the system starting state as an objective function to be optimized in terms of the autonomy levels. A natural result of the construction of the objective function in this way being that the lowest cost ρ thus identified was a local minimum expected cost autonomy level plan for T_G. In this section, we will analyze this algorithm to identify certain conditions under which this local optimum is also a global optimum. This analysis then allows us to identify the conditions under which the joint plans are robustly optimal.

For the purpose of clarity, we construct a permutation $i \rightarrow j$ which re-indexes the subtasks such that they are labeled in the ordering by ρ. Using this ordering, we can write out a proxy function for the cost of a plan:

$$\begin{aligned} \mu_\rho(A_L) = \mu\left(\rho_{A_L}\right) = \left(\mu(A_1) \cdot \mu(A_2) \cdot \ldots \mu\left(A_{A_L}\right)\right) \\ \cdot \left(\mu\left(M_{A_L+1}\right) \cdot \mu\left(M_{A_L+2}\right) \cdot \ldots \mu(M_K)\right) \end{aligned} \tag{5}$$

Because of conditions 1 through 3, this function will possess the same differential landscape as $\Delta E(S)$. We define this function of manual vs. autonomous module cost, f_H, by plotting points $(\mu(A_j),\mu(M_j))$: $f_H(\mu(Aj)) = \mu(Mj)$.

We can further define an uncertainty function $\Delta\mu(M_j)$, to represent the variance in the costs of the manual module j relative to the autonomous modules. Convolving this uncertainty with f_H calculates a statistical model for the cost over $\mu_\rho(A_L)$ as a function of A_L:

$$\begin{aligned} ln\left(\mu_\rho(A_L)\right) = \sum_{j=0}^{A_L} ln\left(\mu\left(A_j\right)\right) + \\ \int_{A_L+1}^{K} \sum_{l=A_L+1}^{K} \delta\mu\left(M_j\right) ln\left(f_H\left(\mu\left(A_l\right)\right)\right) dj \end{aligned} \tag{6}$$

This allows us to use the derivative operator to examine cost-change behavior of μ_ρ in terms of A_L. Applying the operator to both sides and re-arranging the terms, the derivative of the cost function μ evaluated at the autonomous subtask A_L. Seeking extrema, we can set $\mu'_\rho = 0$ to obtain:

$$\mu\left(A_{A_L}\right) ln\left(f_H\left(\mu\left(A_L\right)\right)\right) \int_{A_L+1}^{K} \delta\mu\left(M_{A_L}\right) dj = \mu'\left(A_{A_L}\right) \tag{7}$$

Equation 7 constitutes an implicit relationship describing the rate of change of the cost of an autonomy level plan as a function of A_L. In essence, this equation quantifies relative advantage in our model, and thus naturally mirrors the analysis in (Aghion, 2017).

Though a monotonic relationship between $\mu(A_j)$ and $\mu(M_j)$ will allow for direct optimal autonomy level plan identification, it is fairly unreasonable to expect such regularity of real-world systems. However, Equation 9 is a continuous optimization over cost but the actual autonomy level plan itself

is a set of discrete categories. This suggests the possibility that a measure of deviation from strict monotone behavior may be tolerable,if changes to the function do not result in misidentification of which subtasks should be performed autonomously.

Suppose we can select a monotonic envelope function e_f, as illustrated in Figure 1 and define the deviation for each subtask P_j as r_j, the scale factor required to transform the value of f_H at a point to the corresponding point on e_f: $e_f(i) = r_j\, f_H(i)$.

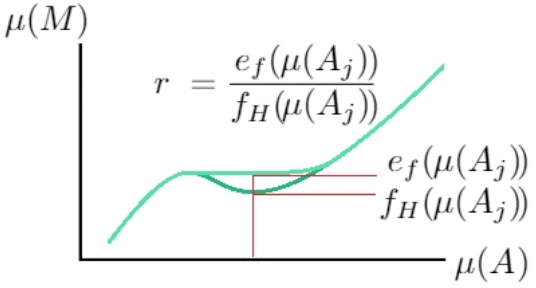

Figure 1: Example envelope function for f_H.

This transform introduces an inaccuracy in the cost function, and we can determine conditions under which the modification causes misclassification of individual modules.

Given $e_f(j) = r_j f_H(\mu(A_j))$, the change in cost for task j is given by $(1–r_j)\mu(Aj)$. For the subtasks which are affected by this change, we have a 'new' autonomy level plan, ρ', determined from e_f rather than f_H with cost and cost changes given by:

$$\Delta\mu = (1 - r_j) \cdot \mu(\rho)^2 \prod_{l=0}^{A_L} \mu\,(A_l) \tag{8}$$

$\mu(\rho') = r_j\mu(\rho)$, and $\Delta\mu=(1–r_j)\mu(\rho)$. We can estimate the expected change in cost as:

$$(1 - r_j) \cdot \mu(\rho)^2 \prod_{l=0}^{A_L} \mu\,(M_l)^{-1} < r_j \cdot \mu(\rho) \tag{9}$$

Which produces a condition under which the change does not lead to misclassification:Then, for multiple deviations from monotonicity with change ratios $r_{j1}, r_{j2}, \ldots r_{js}$, the condition for optimality is given by:

$$\prod_{l=0}^{s} \frac{1 - r_{jl}}{r_{jl}} < \left(\frac{\prod_{l=1}^{A_L} \mu(M_j)}{\mu(\rho)} \right)^s \tag{10}$$

EXPERIMENTS

To experimentally validate the predictions of our analysis, we have applied it to optimization of Human/Robot interaction on a physical robot performing a sanitization task. We collected operational data comprising subtask costs, in terms of time and the probabilities of successful completion of each subtask autonomously and with 25 human operators.

We decompose our disinfection task into seven sub-tasks: Target Identification, Alignment of robot to target, Corrections to object tracking, Assignment of kinematic bounds, Measurement and tracking mode selection, Placement of end-effector, and Tracking of coverage trajectory. The data are collected throughout each trial by recording the actions of the user via UI scripting. Time costs are measured as simple duration from beginning one task to completing it. Failure chance is measured in two ways- first, all instances in which the robot system encounters a significant fault, such as a tracking excursion in which the target is fully lost. For human execution, cases where the operator must re-try are marked as a failure state for that specific subtask module.

Table 1: Average performance metrics.

	Autonomous			Manual		
Module	**Prob.**	**Time**	σ	**Prob.**	**Time**	σ
Target Id	0.90	24s	0.27	0.37	36s	5.7
Alignment	0.50	17s	0.25	0.32	35s	3.8
Object track	0.41	13s	0.22	0.16	49s	1.6
Kinematics	0.67	15 s	0.15	0.23	47 s	1.5
Tracking	0.36	24s	0.12	0.15	13s	2.5
Placement	0.16	11 s	0.10	0.27	15 s	0.9
Coverage	0.13	21 s	0.08	0.15	13s	2.6

Table 2: Derived typical autonomy level plans.

A_L / Costs	1	2	3	4	5	6	7
Average (all)	69.5	56.3	47.1	42.2	42.3	44.7	50.7
Validation Trials	59.1	54.6	49.8	46.5	50.2	45.2	50.7
Lowest Performer	78.7	56.9	48.3	48.5	49.0	49.9	50.7
% diff.	15%	3%	6%	10%	19%	1%	-

Table 1 shows the probabilities and execution times for each of the subtasks averaging over all samples, along with the variance for each. On Table 2, we see that the autonomy levels determined have a minimum cost of 42.2 at AL = 4. Also shown is a low-performing user with an optimum at AL = 3. We solve an MDP implementation of this task with the reward function defined by the time cost of each step, $Ra(s,s') = -\Delta t$. With this method, we find an optimum plan cost of 42.67s, a 1.1% error with respect to the minimum time predicted. We can also, critically, use the data collected in

our experiments to validate the robustness analysis. We construct estimates of fH(j) by fitting curves, seen on Figure 1. Table 3 presents the details of of Equations 9 & 10 for the average case and a low-performing user. In the average case, we see that the equation has the minimum difference 0.12, at AL = 4, identical to the prediction. Likewise, for the low-performer, the nearest correspondence is at the determined optima AL = 3. This shows that for both cases, the measured values in Equation 10 hold at the predicted points.

Table 3: Applying theorem I.

	Average Case		Low Performer	
μ'	$\mu\ ln\left(f_H\right) \int \delta\mu$	Δ	$\mu ln\left(f_H\right) \int \delta\mu$	Δ
1.33	0.14	1.18	14.93	16.26
1.71	0.26	1.44	11.26	12.97
2.09	0.82	1.26	0.25	1.83
2.47	2.34	0.12	0.56	1.90
2.85	1.31	1.53	0.59	2.25
3.23	1.08	2.14	0.44	2.78
3.61	1.44	2.16	0.62	2.98

Table 4: Applying theorem II.

	Average Case		Low Performer	
$\mu\left(M_j\right)$	$\left(1-r_j\right)/r_j$	$\Pi\mu(M/P)$	$\left(1-r_j\right)/r_j$	$\Pi\mu(M/P)$
2	0.0014	0.021	0.056	0.035
5	-	0.083	-	0.121
7.8	-	0.187	0.711	0.255
8.4	0.437	0.294	0.696	0.400
9	-	0.408	-	0.555
15	0.151	0.597	-	0.814
18	-	0.825	3.672	1.124
	$9.557 \cdot 10^{-5} < 0.562$		$0.102 < 1.597$	

These curves are not strictly monotonic, as expected in real-world situations, and thus present the chance to validate Equation 10. For each subtask, we calculate the change ratio necessary to bring the fit function into compliance with a monotone function. For those modules which are non-conforming, the last row of Table 4 contains (1–r)(r–1), and the corresponding ratio of autonomous to autonomy level plan cost. These values show that in both the general and single-user case the robustness condition is met by at least an order of magnitude.

CONCLUSION

In this paper, we sought to demonstrate with robust mathematical rigor the advantages of Human/AI teams in efficacy over that of homogeneous teams of only humans or only AIs. This work builds on the results referenced in the

Introduction, in which many cases of superior performance was observed in such teams- performance we sought to validate with an analytical framework.

Towards this end, we constructed a generalized model for joint task accomplishment which frames the problem as assignment of subtasks rather than seeking individual optimized plans. Using this analysis, we parametrize the impact of relative advantage between the human and the AI, and plan along a convex, optimizeable surface. Most importantly, we are able to show that the conditions underwriting this optimization remit to robust bounds which rely on the discrete choice in a task being performed by the human versus the AI.

The key and critical insight from this analysis is that the condition expressed in Equation 10 is of a geometrically decreasing nature in the number of subtask steps- that is to say, the range of deviation from ideal conditions widens as task complexity increases. This agrees with our early observation that abstract tasks revealed the mechanisms by which joint systems achieve superior potency. We thus infer that as a task becomes increasingly complex, the probability that the solution of an arbitrary problem is optimal at a joint plan increases geometrically.

These predictions are mirrored in the referenced works, however, we validate them by examining a specific constructed task with the express purpose of collecting the data required to examine all the quantities defined within our framework. We find that these measurements match with our predictions, and that the bounds thus established hold, to within an order of magnitude of clearance.

We therefor draw the conclusion that not only is it an observable emergent property that joint teams often out-perform unitary teams, but in fact a natural conclusion of the mechanism of relative advantage between them. The relative impact of cognitive loading is one such discrepancy, but as we have seen the nature of that difference is less important than the difference itself. The fact of existence of distinctions between a human mind and an AI system itself guarantees, as is often observed within economic systems, that a combination of contributions will naturally effect the strongest results. Humans and AIs working in tandem are more effective for the same reasons that international trade is more profitable than autarky.

REFERENCES

Aghion, P., Jones, B. F., & Jones, C. I. (2017). Artificial intelligence and economic growth (Vol. 23928). Cambridge, MA: National Bureau of Economic Research.

Barber, K. S., & Martin, C. E. (1999, May). Agent autonomy: Specification, measurement, and dynamic adjustment. In Proceedings of the autonomy control software workshop at autonomous agents (Vol. 1999, pp. 8–15).

Crandall, J. W., & Goodrich, M. A. (2002, September). Characterizing efficiency of human robot interaction: A case study of shared-control teleoperation. In IEEE/RSJ international conference on intelligent robots and systems (Vol. 2, pp. 1290–1295). IEEE.

Feng, L., Wiltsche, C., Humphrey, L., & Topcu, U. (2016). Synthesis of human-in-the-loop control protocols for autonomous systems. IEEE Transactions on Automation Science and Engineering, 13(2), 450–462.

Fu, J., & Topcu, U. (2015). Synthesis of shared autonomy policies with temporal logic specifications. IEEE Transactions on Automation Science and Engineering, 13(1), 7–17.

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. Computers in Human Behavior, 139, 107502.

Kim, D. J., Hazlett-Knudsen, R., Culver-Godfrey, H., Rucks, G., Cunningham, T., Portee, D.,... & Behal, A. (2011). How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 42(1), 2–14.

Li, Y., Li, Y., Chen, Q., & Chang, Y. (2024). Humans as teammates: The signal of human–AI teaming enhances consumer acceptance of chatbots. International Journal of Information Management, 76, 102771.

Memmert, L., & Tavanapour, N. (2023). Towards human-AI-collaboration in brainstorming: Empirical insights into the perception of working with a generative AI.

Nichols, K. A., & Okamura, A. M. (2015). A framework for multilateral manipulation in surgical tasks. IEEE Transactions on automation science and engineering, 13(1), 68–77.

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, 5(1), 46–57.

Ulfert, A. S., Georganta, E., Centeio Jorge, C., Mehrotra, S., & Tielman, M. (2024). Shaping a multidisciplinary understanding of team trust in human-AI teams: A theoretical framework. European Journal of Work and Organizational Psychology, 33(2), 158–171.

Vassilakopoulou, P., Haug, A., Salvesen, L. M., & Pappas, I. O. (2023). Developing human/AI interactions for chat-based customer services: Lessons learned from the Norwegian government. European journal of information systems, 32(1), 10–22.

Vold, K. (2024). Human-AI cognitive teaming: Using AI to support state-level decision making on the resort to force. Australian Journal of International Affairs, 78(2), 229–236.

Westby, S., & Riedl, C. (2023, June). Collective intelligence in human-AI teams: A Bayesian theory of mind approach. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 5, pp. 6119–6127).