Designing Algorithmic Delegates

Sophie Greenwood
Cornell Tech
sjgreenwood@cs.cornell.edu

Karen Levy
Cornell University
karen.levy@cornell.edu

Solon Barocas
Microsoft Research
solon@microsoft.com

Jon Kleinberg
Cornell University
kleinberg@cornell.edu

Hoda Heidari
Carnegie Mellon University
hheidari@cmu.edu

Abstract

As AI technologies improve, people are increasingly willing to *delegate* tasks to algorithmic agents. A human decision-maker decides whether to delegate to an AI agent based on features of the decision-making instance they are faced with; since humans typically lack full awareness of these features, they perform a kind of *categorization* by treating decision-making instances that agree in all their observable features as indistinguishable from one another. In this paper, we define the problem of designing the optimal algorithmic delegate in the presence of categorization, and reveal the fundamentally combinatorial nature of this problem. We show that finding the optimal delegate is computationally hard in general, but we find an efficient algorithm for a large family of settings.

1 Introduction

Algorithmic agents – such as robots, AI assistants, and chatbots – are increasingly accurate and effective. In many cases, a human user works with the algorithmic agent to select an action, and the quality of the algorithmic agent must be measured by the performance of the combined human-algorithm system, or the *team performance* [1, 2]. Recent work has demonstrated that the algorithmic agent with the highest standalone accuracy does not necessarily achieve the optimal team performance [2, 3]. Thus, it is important and natural to ask: *How can one design the optimal algorithmic teammate?*

The answer to this depends on the nature of the team. There are many ways a human may partner with an algorithmic agent; we focus on teams where the machine is a *delegate* to whom the human may hand off the choice of action [4–6]. In cases where it is impossible or prohibitively time-consuming for the human and algorithm to communicate before selecting each action, the algorithm must be a delegate. Moreover, as AI technology improves, users may prefer to delegate low-stakes decisions to increase efficiency. Examples of algorithmic delegates include autonomous vehicles where a driver must decide whether to drive herself or to engage an AI system [7, 8], or AI assistants that may be dispatched to answer emails, schedule meetings, or buy products online [9, 10].

A key question is whether a human user – presented with a scenario where an action must be selected and taken – will delegate the choice of action or not. In real scenarios, the human will not have complete information about the specific instance of the decision-making problem she faces, and as a result cannot make this delegation decision perfectly [1]. For example, the user of an AI shopping agent may not have details about past trends in the online market, and a driver may not know about rough road conditions ahead or a nearby hidden driveway that tends to cause accidents. The human must then make the best delegation and action decisions she can given what she can observe. We refer to these sets of human-indistinguishable instances as *categories*. The algorithmic agent – hereafter, the *machine* – similarly has incomplete information [11]. The AI shopping agent may not have full

NeurIPS 2024 Workshop on Behavioral Machine Learning.

knowledge of the user's preferences, and the autonomous vehicle may not know about a large sports event in the neighborhood and other social context. Thus the machine also only observes its own *categories* of states.

Overview of results. In this work, we investigate how to design the optimal algorithmic delegate in the presence of categorization. We first develop a simple model of human delegation to an algorithmic agent where both the human and algorithm have incomplete information about the world, and set up the problem of finding the optimal algorithmic delegate. We prove a characterization theorem which says the optimal delegate is designed for the categories in which it is used, and reveals a fundamental combinatorial structure in the problem. We then show that finding the optimal algorithmic delegate is tractable when the ground truth optimal action is a decomposable function of the human and machine features, but that it is NP-hard in general.

Related work. This paper extends a long line of work addressing how to design algorithmic agents to improve performance in human-AI teams [2, 12, 6, 13, 14]. In particular, we focus on cases in which the human must choose to either delegate to the machine [5, 15] or take action by herself; to our knowledge, we are the first to study the design of optimal delegates that account for categorization. Bansal et al. [2] propose that a machine's performance should be optimized for the categories where it is used, but do not study delegation. Other work studies machine design in other team structures, such as settings where the machine delegates to the human [16, 17], settings where the human may observe the machine's output before deciding [2, 12], and settings where a human chooses between accepting a machine's prediction and delegating to another human [18].

Categorization is a well-known aspect of human decision-making [19, 20]. This act of agents grouping decision-making scenarios has been modeled both as a behavioral phenomenon [19–21], and as a consequence of limited information [1, 11, 22]. We focus on the second type of categorization; our focus on information-driven categorization is consistent with contemporary human-algorithm collaboration work: "mental models" [1] and "indistinguishable" inputs [11] are analogous to our human and machine categories respectively. Moreover, our human categories are captured by "generalization functions" in Vafa et al. [22] when the human maintains equivalent beliefs about all states with the same human-accessible features. Iakovlev and Liang [23] also study a model where a human and machine have access to different information in the form of binary features, but focus on the problem of a third party selecting an evaluator. Some of these prior works study delegation in the presence of categorization [5, 22], but do not investigate the question of optimal delegate design.

This work also interfaces with a variety of other disciplines. Our model can represent a type of *interpretability* [24] by the number of shared features between the human and machine: in our model, more shared features corresponds to more similar categories and actions taken. Categorization can cause *over-reliance* [25]: the need to make the same delegation decision across a category can result in over- or under-delegation within categories. In human factors analysis, function allocation prescribes a qualitative process by which a designer determines which tasks in a system should be automated [26, 27]; we take a quantitative approach which is more suited to modern AI design.

2 Model

Let the world be described by d binary features $x_1, ..., x_d \in \{0, 1\}$. There are then $n = 2^d$ possible *states* of the world $\mathbf{x} = (x_1, ..., x_d) \in \{0, 1\}^d$; for simplicity, we let each state occur with equal probability. In each state $\mathbf{x}$, there is some ground truth correct action $f^*(\mathbf{x}) \in \mathbb{R}$ that any agent should take. If an agent takes action $a \in \mathbb{R}$, it will receive a loss of $(a - f^*(\mathbf{x}))^2$.

There are two agents, a human and a machine. The human can observe features $I_H \subseteq \{1, ..., d\}$; the machine can observe features $I_M \subseteq \{1, ..., d\}$. A *delegation setting* is determined by I_H, I_M, and f^*. While I_H and I_M can each be arbitrary, we will assume for simplicity that the set of features is partitioned into I_H and I_M; in Appendix C we show how our results generalize. Let $\mathbf{x}_H$ and $\mathbf{x}_M$ denote the restrictions of $\mathbf{x}$ to human and machine-observable features.

To situate how this formalism works through a brief example, suppose that the machine is an AI shopping agent as in the introduction, which can traverse the Web to purchase items on a user's behalf. In this case, $\mathbf{x}$ represents the features of an item on a given day, and the ground truth optimal action $f^*(\mathbf{x})$ is the highest price a user should pay for that item on that day. The human-observable features I_H could be the users' preferences and level of urgency for purchasing the item, and I_M

could be information on how the item's current price compares to market trends for that item on the Web. The human and machine might share some features – the human could communicate some preferences to the machine – but the human cannot fully articulate the subtleties of her preferences, (such as how urgently she needs an item or her preferences for substitutes) and the machine cannot fully summarize all the complex market trends involved in a way that's legible to the human.

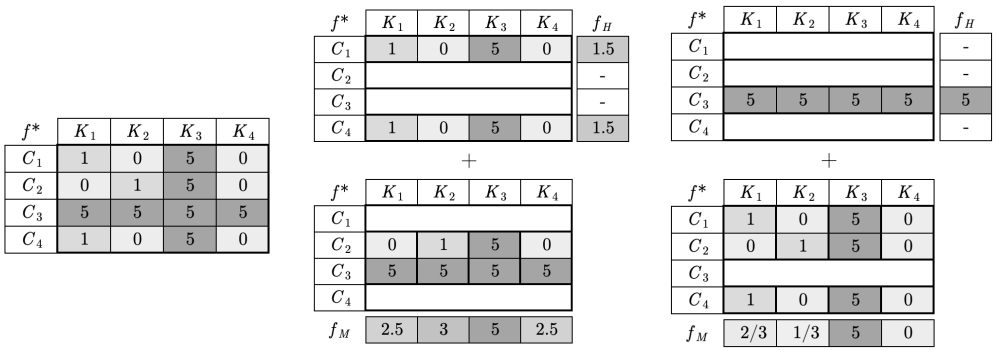
A human category C is a set of states that are indistinguishable to a human because they all share the same human-observable features. Formally, states $\mathbf{x}$ and $\mathbf{z}$ are in the same human category if and only if $\mathbf{x}_H = \mathbf{z}_H$. Similarly, a machine category K is a set of states indistinguishable to the machine, so that states $\mathbf{x}$ and $\mathbf{z}$ are in the same machine category if and only if $\mathbf{x}_M = \mathbf{z}_M$. Let $\mathcal{H}$ and $\mathcal{M}$ denote the set of all human and machine categories respectively; $\mathcal{H}$ and $\mathcal{M}$ are each a partition of the states. We can enumerate the human categories $C_1, C_2, ..., C_h \in \mathcal{H}$ and the machine categories $K_1, K_2, ..., K_m \in \mathcal{M}$. For a state $\mathbf{x}$, let $C(\mathbf{x})$ and $K(\mathbf{x})$ be the human and machine categories containing $\mathbf{x}$. Note that when I_H and I_M partition the feature set there is a single state $\mathbf{x}_{ij} \in C_i \cap K_j$.

Since each agent can't distinguish between states within a category, the human and machine choose actions as a function of the category they observe. Let $f_H : \mathcal{H} \to \mathbb{R}$ and $f_M : \mathcal{M} \to \mathbb{R}$ denote the human and machine's action functions. The delegation process works as follows.

1. Given state $\mathbf{x}$, the human observes the category $C(\mathbf{x})$.
2. The human decides whether or not to delegate based on which agent has better expected performance in $C(\mathbf{x})$.
3. If the human does not delegate, she takes action $f_H(C(\mathbf{x}))$.
4. If the human delegates, the machine observes the machine category $K(\mathbf{x})$, and takes action $f_M(K(\mathbf{x}))$.

The machine's designer – typically another human – is aware of this process, and must design f_M accordingly; we investigate the optimal design for f_M. Note that we assume the human can observe both her own and the machine's expected loss in each category; for example she may learn this over time through interactions with the machine [28].

Returning to the shopping agent example, the human can observe her preferences but not the market trends, resulting in a set of indistinguishable states corresponding to a human category C. Based on C, she decides whether to delegate to the machine, or take action $f_H(C)$. If she delegates to the machine, the machine observes the market trends but not all aspects of the user's preferences, resulting in a set of indistinguishable states that corresponds to a machine category K, and takes action $f_M(K)$. Our results address how a designer should create the optimal agent f_M for this setting.



(a) Ground truth correct actions (b) Delegate in $\{C_2, C_3\}$ (c) Delegate in $\{C_1, C_2, C_4\}$

Figure 1: Example delegation problem.

An illustrative example. Before formalizing the general delegation problem, we will work through the example in Figure 1. Here, the human and machine each have four categories (as would arise when each has access to two underlying Boolean features). Figure 1a visualizes these categories in a

grid, where the rows are human categories and the columns are machine categories. The value of entry (i, j) is the ground truth correct action $f^*(\mathbf{x}_{ij})$.

In a given state $\mathbf{x}_{ij}$, the human can observe which row i the state is in, and will either delegate to the machine, or take action $f_H(C_i)$. Suppose that the human decides to delegate in categories $\{C_2, C_3\}$. As illustrated in the upper grid of Figure 1b, her loss-minimizing action in categories C_1 and C_4 is to take the average action across the category. When the human delegates, the machine can observe the column j and take some action $f_M(K_j)$. If the machine knows that the human doesn't delegate in states C_1 and C_4, its loss-minimizing action is to take the average action across only C_2 and C_3, shown in the lower grid of Figure 1b.

However, these f_H and f_M have very high loss. Can we do better? In this example, states in the human category C_3 are difficult for the machine to handle: the optimal action 5 is often very different than in other states in the machine's categories. However, C_3 is very easy for the human to take action in: the human can take action 5 with no loss, as illustrated in the top grid in Figure 1c. Without C_3, the remaining values in each machine category have very low variance, and the machine achieves very good performance even if the human delegates in all other human categories. Indeed, we can show that the f_M shown in Figure 1c that averages over the human categories excluding C_3 is optimal.

The example above suggests two high-level insights. First, the optimal delegate partitions human categories into "delegate" and "non-delegate" categories, is designed to only be used in "delegate" categories, and is in fact only used in "delegate" categories. We generalize this in Proposition 1. Second, to find this partition, the machine's designer should take the non-delegate human categories to be those with low variance, and the delegate human categories to be those that make the machine categories have low variance. The optimal delegate is then designed to only be used in the latter set of categories. In Proposition 2 we show this is exactly the problem of designing the optimal delegate.

Team loss minimization. We now formalize the general delegation problem. Let $\ell_H(f_H, C)$ and $\ell_M(f_M, C)$ denote the expected loss of the human function f_H and machine function f_M respectively in category C. Recall that we assume that the human can observe the machine's average loss in a category $\ell_M(f_M, C)$. When the human is delegating optimally, she will delegate to the machine in category C if and only if the machine's loss is lower than the human's in that category. The loss associated with the team, in which the human with function f_H optimally delegates tasks to the machine with function f_M, is denoted by $\ell(f_H, f_M)$. For simplicity, we refer to this as the *team loss*. Formally, the expected team loss is $\ell(f_H, f_M) := \frac{1}{|\mathcal{H}|} \sum_{C \in \mathcal{H}} \min \{\ell_H(f_H, C), \ell_M(f_M, C)\}$.

In a given category C, the human will observe C, and either delegate to the machine, or will take an action. If the human chooses to take an action, given that the human can only observe C, the loss-minimizing action in category C is the average value of f^* in C; we denote the human function that takes this optimal action in each C by f_H^*. This will be the optimal human function regardless of the machine's function f_M, so we will restrict our consideration to $f_H = f_H^*$.

The optimal machine action is less clear: in machine category K, the machine has access not only to the features corresponding to K, but also to the information that the human chose to delegate to it (by nature of having to choose an action at all). The machine could be oblivious, relying only on the machine-interpretable features and ignoring the fact that it has been delegated to. In this case the optimal action in machine category K would be the average value of f^* across states in K; we denote this function by f_M^{obliv}. Alternatively, an *optimal delegate* f_M^* makes use of all the information available and attains the optimal team loss, $f_M^* \in \arg\min_{f_M} \ell(f_H^*, f_M)$.

3 Results

Our goal is now to find optimal delegates f_M^*. We first state two results (Propositions 1 and 2) that give general versions of the principles from our example above. We then use these results to show that we can efficiently find the optimal delegate for a large family of functions f^* (Theorem 3), but that this is hard in general (Theorem 5). We provide proofs in Appendix B.

Reformulating the problem. In Proposition 1, we transform this problem into a discrete optimization problem, and, in Proposition 2, we show this discrete problem has a novel combinatorial formulation.

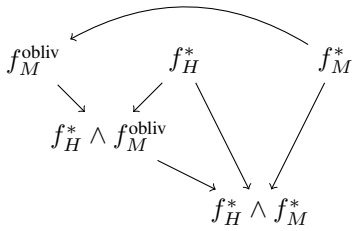


Figure 2: Relationships between the losses of different possible human and machine teams. $f_H \wedge f_M$ denotes a human with function f_H optimally delegating to the machine f_M; an arrow from team A to team B indicates that the loss of team B will always be (weakly) lower than team A. See Appendix A for detailed explanations.

Proposition 1 (Informal)**.** *To find an optimal delegate f_M^*, it is sufficient and necessary to find a set of human categories $\mathcal{R}$ that attains the minimum team loss when the human delegates to $f_M^{\mathcal{R}}$ in precisely the categories in $\mathcal{R}$.*

Proposition 2. *Define a matrix $V \in \mathbb{R}^{|\mathcal{H}|} \times \mathbb{R}^{|\mathcal{M}|}$ with entries $v_{ij} = f^*(\mathbf{x}_{ij})$. The problem of finding an optimal delegate is as follows:*

> VARIANCEASSIGNMENT. Fix a set of rows S of V. For each row $i \in S$, pay a cost proportional to the variance of V across row i, and remove row i from V. Then, for each column j, pay a cost proportional to the variance across column j of the remaining entries. Find a set S^* that minimizes the total cost.

Then for $\mathcal{R} = \{C_i : i \notin S^*\}$, $f_M^{\mathcal{R}}$ will be an optimal delegate.

Tractability. We first show that if f^* is separable, that is, f^* can be decomposed additively into functions of the human and machine features respectively, we can efficiently find the optimal delegate. Note that linear functions are separable.

Theorem 3. *Suppose that f^* is separable, that is, $f^*(\mathbf{x}) = u(C(\mathbf{x})) + w(K(\mathbf{x}))$ for some functions u, w. Then we can find an optimal delegate f_M^* in time polynomial in the size of f^*.*

We also find that if the human or machine has access to a limited number of features, the problem is again tractable.

Theorem 4. *Suppose that $|I_H| = O(1)$ or $|I_M| = O(1)$. Then we may find an optimal delegate in time polynomial in the size of f^*.*

However, there is no general efficient algorithm to find the optimal delegate.

Theorem 5. *Unless* P = NP*, there is no algorithm to find an optimal delegate f_M^* in time polynomial in the size of f^* for all ground truth functions f^*.*

This result motivates why the optimal delegate has not previously been characterized: the problem has a fundamentally combinatorial nature that makes it intractable to solve.

4 Discussion

In this paper, we developed a formal model for settings where a human decides whether to delegate to a machine, and showed that categorization arises from information asymmetries. There are many different machine designs, ranging from an *oblivious* machine that ignores that it is a delegate, to an *optimal* machine that accounts for delegation. In Figure 2 we describe general relationships in performance between some of these machines. We studied the problem of designing an optimal machine for delegation in the presence of categorization. We showed that this induces surprisingly clean algorithmic formulations, and derived tractability results.

An interesting direction for future work is determining other delegation settings in which finding an optimal delegate is tractable. Moreover, investigating which ground truth optimal action functions f^* lead to qualitatively different optimal delegates could yield heuristics for intractable instances.

Acknowledgments and Disclosure of Funding

The authors would like to thank Erica Chiang, Kate Donahue, Greg d'Eon, Malte Jung, and Kenny Peng for valuable discussions and feedback, as well as the Artificial Intelligence, Policy, and Practice (AIPP) group at Cornell. The authors also thank the NeurIPS 2024 Workshop on Behavioral Machine Learning reviewers for helpful suggestions.

SG is supported by a fellowship from the Cornell University Department of Computer Science and an NSERC PGS-D fellowship [587665]. The work is supported in part by a grant from the John D. and Catherine T. MacArthur Foundation.

References

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):2–11, Oct. 2019. doi: 10.1609/hcomp.v7i1.5285. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. Is the most accurate AI the best teammate? Optimizing AI for teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11405–11414, May 2021. doi: 10.1609/aaai.v35i13.17359. URL https://ojs.aaai.org/index.php/AAAI/article/view/17359.

[3] Karim Hamade, Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Designing skill-compatible AI: Methodologies and frameworks in chess. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Nathan Stout, Alan Dennis, and Taylor Wells. The buck stops there: The impact of perceived accountability and control on the intention to delegate to software agents. *AIS Transactions on Human-Computer Interaction*, 6, 03 2014. doi: 10.17705/1thci.00058.

[5] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3501999. URL https://doi.org/10.1145/3491102.3501999.

[6] Allen E. Milewski and Steven H. Lewis. Delegating to software agents. *International Journal of Human-Computer Studies*, 46(4):485–500, 1997. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.1996.0100. URL https://www.sciencedirect.com/science/article/pii/S1071581996901007.

[7] General Motors. Hands-free, eyes on, 2024. URL https://www.gm.com/innovation/av-safe-deployment.

[8] Tesla. Autopilot and full self-driving (supervised), 2024. URL https://www.tesla.com/support/autopilot.

[9] Booked AI. How AI travel can curate your perfect wine tour in Italy, 2024. URL https://www.booked.ai/blogs/how-ai-travel-can-curate-your-perfect-wine-tour-in-italy.

[10] Wing Assistant. What can Wing General VAs do?, 2024. URL https://wingassistant.com.

[11] Rohan Alur, Manish Raghavan, and Devavrat Shah. Human expertise in algorithmic prediction, 2024. URL https://arxiv.org/abs/2402.00793.

[12] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2429–2437, Jul. 2019. doi: 10.1609/aaai.v33i01.33012429. URL https://ojs.aaai.org/index.php/AAAI/article/view/4087.

[13] Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-AI complementarity with predictions sets, 2024. URL https://arxiv.org/abs/2405.17544.

[14] Nina Grgić-Hlača, Claude Castelluccia, and Krishna P. Gummadi. Taking advice from (dis)similar machines: The impact of human-machine similarity on machine-assisted decision-making. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):74–88, Oct. 2022. doi: 10.1609/hcomp.v10i1.21989. URL https://ojs.aaai.org/index.php/HCOMP/article/view/21989.

[15] Brian Lubars and Chenhao Tan. Ask not what AI can do, but what AI should do: towards a framework of task delegability. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[16] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. 2018.

[17] Thodoris Lykouris and Wentao Weng. Learning to defer in content moderation: The human-AI interplay, 2024. URL https://arxiv.org/abs/2402.12237.

[18] Ruqing Xu. Persuasion, delegation, and private information in algorithm-assisted decisions, 2024. URL https://arxiv.org/abs/2402.09384.

[19] John R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3): 409–429, 1991. doi: 10.1037/0033-295x.98.3.409.

[20] Sendhil Mullainathan. Thinking through categories, 2002.

[21] Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. Concept alignment, 2024. URL https://arxiv.org/abs/2401.08672.

[22] Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. Do large language models perform the way people expect? Measuring the human generalization function. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48919–48937. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/vafa24a.html.

[23] Andrei Iakovlev and Annie Liang. The value of context: Human versus black box evaluators, 2024. URL https://arxiv.org/abs/2402.11157.

[24] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832. URL https://www.mdpi.com/2079-9292/8/8/832.

[25] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI: Literature review. Technical Report MSR-TR-2022-12, Microsoft, June 2022. URL https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

[26] Philip Marsden and Mark Kirby. Allocation of functions. *Handbook of human factors and ergonomics methods*, pages 34–1, 2005.

[27] Harold E. Price. The allocation of functions in systems. *Human Factors*, 27(1):33–45, 1985. doi: 10.1177/001872088502700104. URL https://doi.org/10.1177/001872088502700104.

[28] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M. Drucker. What did my AI learn? How data scientists make sense of model behavior. *ACM Trans. Comput.-Hum. Interact.*, 30(1), Mar 2023. ISSN 1073-0516. doi: 10.1145/3542921. URL https://doi.org/10.1145/3542921.

[29] YXD. Computing the subset giving the minimum standard deviation in an array. *Stack Overflow*, 2013. URL https://stackoverflow.com/a/20143904.

[30] B. O'Neill. Some useful moment results in sampling problems. *The American Statistician*, 68 (4):282–296, 2014. ISSN 00031305. URL http://www.jstor.org/stable/24591747.

[31] Bernard Chazelle, Herbert Edelsbrunner, Leonidas J. Guibas, and Micha Sharir. A singly exponential stratification scheme for real semi-algebraic varieties and its applications. *Theoretical Computer Science*, 84(1):77–105, 1991. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(91)90261-Y. URL https://www.sciencedirect.com/science/article/pii/030439759190261Y.

[32] Ulrik Brandes, Eugenia Holm, and Andreas Karrenbauer. Cliques in regular graphs and the core-periphery problem in social networks. In T-H. Hubert Chan, Minming Li, and Lusheng Wang, editors, *Combinatorial Optimization and Applications*, pages 175–186, Cham, 2016. Springer International Publishing. ISBN 978-3-319-48749-6.

[33] Charalampos E. Tsourakakis, Tianyi Chen, Naonori Kakimura, and Jakub Pachocki. Novel dense subgraph discovery primitives: Risk aversion and exclusion queries. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, page 378–394, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-46149-2. doi: 10.1007/978-3-030-46150-8_23. URL https://doi.org/10.1007/978-3-030-46150-8_23.

A Loss comparisons

In this section, we give justifications for the relationships captured in Figure 2. Define $\ell(f_A)$ as the loss of agent A acting individually.

First, any model with the human optimally delegating between f_H and some machine f_M will always be better than f_H and f_M acting individually, because the human could choose to always delegate, thus emulating f_M, or never delegate to emulate f_H. Formally, $\ell(f_H, f_M) \leq \ell(f_H)$ and $\ell(f_H, f_M) \leq \ell(f_M)$. This produces the arrows from the standalone models to the model teams.

Moreover, $\ell(f_H^*, f_M^*) \leq \ell(f_H^*, f_M)$ for any machine f_M. This implies that delegating to the oblivious machine will have worse loss than delegating to the optimal delegate f_M^*, i.e., $\ell(f_H^*, f_M^*) \leq \ell(f_H^*, f_M^{\text{obliv}})$. This results in an arrow from $f_H^* \wedge f_M^{\text{obliv}}$ to $f_H^* \wedge f_M^*$.

However, without delegation, the oblivious machine will perform better than the optimal delegate, $\ell(f_M^{\text{obliv}}) \leq \ell(f_M^*)$: the oblivious machine is defined to be the best machine without delegation. This results in the final arrow from f_M^* to f_M^{obliv}.

The last two relationships we described echo the results of [2] and [3], who show in other human-algorithm collaboration settings the optimal individual algorithmic agent is a worse collaborator than an algorithmic agent designed for collaboration.

B Theorem statements and proofs

In this appendix, we formalize and prove our theoretical results.

B.1 Formal statement and proof of Proposition 1

First, we formally define

$$f_M^{\mathcal{R}}(K) := \frac{1}{\sum_{C \in \mathcal{R}} |K \cap C|} \sum_{C \in \mathcal{R}} \sum_{\mathbf{x} \in K \cap C} f^*(\mathbf{x}).$$

Recall Proposition 1.

Proposition 1 (Informal)**.** *To find an optimal delegate f_M^*, it is sufficient and necessary to find a set of human categories $\mathcal{R}$ that attains the minimum team loss when the human delegates to $f_M^{\mathcal{R}}$ in precisely the categories in $\mathcal{R}$.*

Define $\mathcal{D}(f_H, f_M) = \{C \subseteq \mathcal{H} : \ell_M(f_M, C) < \ell_H(f_H, C)\}$ to be the set of categories in which a human with function f_H will delegate to a machine with function f_M.

We formalize Proposition 1 as follows.

Proposition. *Let f^* be a ground truth function. Recall that an optimal delegate f_M^* is defined as a solution to the optimization problem*

$$\min_{f_M} \ell(f_H^*, f_M) \equiv \min_{f_M} \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{D}(f_H^*, f_M)} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{D}(f_H^*, f_M)} \ell_M(f_M, C) \right). \quad (1)$$

Consider the combinatorial optimization problem

$$\min_{\mathcal{R} \subseteq \mathcal{H}} \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}} \ell_M(f_M^{\mathcal{R}}, C) \right). \quad (2)$$

If $\mathcal{R}$ is a solution to Problem 2, then $f_M^{\mathcal{R}}$ is a solution to Problem 1.

Likewise, if f_M^* is a solution to Problem 1, then $\mathcal{D}(f_H^*, f_M^*)$ is a solution to Problem 2.

Proof. First, notice that Problem 1 can be expressed as

$$\begin{aligned} \min_{f_M, \mathcal{R}} \quad & \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}} \ell_M(f_M, C) \right) \\ \text{s.t.} \quad & \mathcal{R} = \mathcal{D}(f_H^*, f_M) \end{aligned}$$

Moreover, Problem 2 can be expressed as

$$\begin{aligned} \min_{f_M, \mathcal{R}} \quad & \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}} \ell_M(f_M, C) \right) \\ \text{s.t.} \quad & f_M = f_M^{\mathcal{R}} \end{aligned}$$

We define an intermediate problem,

$$\min_{f_M, \mathcal{R}} \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}} \ell_M(f_M, C) \right) \quad (3)$$

We will show that an optimal solution to Problem 3 is feasible for both Problems 1 (Part 1) and 2 (Part 2). Since all three functions have the same objective, this implies that the set of solutions will be the same for each problem.

This implies that if f_M^* solves Problem 1, then $f_M^*, \mathcal{D}(f_H^*, f_M^*)$ solves Problem 3, and thus $\mathcal{D}(f_H^*, f_M^*)$ solves Problem 2. Likewise, if $\mathcal{R}$ solves Problem 2, then $f_M^{\mathcal{R}}, \mathcal{R}$ solves Problem 3, and $f_M^{\mathcal{R}}$ solves Problem 1.

Part 1. Suppose $f_M, \mathcal{R}$ is an optimal solution to Problem 3.

First, for the sake of contradiction suppose that $f_M, \mathcal{R}$ is not a feasible solution to Problem 1.

The constraint that $\mathcal{R} = \mathcal{D}(f_H^*, f_M)$ is equivalent to requiring that both $\ell_M(f_M, C) < \ell_H(f_H^*, C) \implies C \in \mathcal{R}$ and $\ell_H(f_H^*, C) < \ell_M(f_M, C) \implies C \in \mathcal{H} \setminus \mathcal{R}$.

If for some $C \in \mathcal{H} \setminus \mathcal{R}$, $\ell_M(f_M, C) < \ell_H(f_H^*, C)$, let $\mathcal{R}' = \mathcal{R} \cup \{C\}$. Then

$$\begin{aligned} & \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}'} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}'} \ell_M(f_M, C) \right) - \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}} \ell_M(f_M, C) \right) \\ & = \frac{1}{|\mathcal{H}|} \left(\ell_M(f_M, C) - \ell_H(f_H^*, C) \right) \\ & < 0 \end{aligned}$$

and $f_M, \mathcal{R}$ is not an optimal solution to Problem 3. If instead there is some $C \in \mathcal{R}$, and $\ell_H(f_H^*, C) < \ell_M(f_M, C)$ we can similarly let $\mathcal{R}' = \mathcal{R} \setminus \{C\}$, and

$$\begin{aligned}
&\frac{1}{|\mathcal{H}|}\left(\sum_{C\in\mathcal{H}\setminus\mathcal{R}'}\ell_H(f_H^*,C)+\sum_{C\in\mathcal{R}'}\ell_M(f_M,C)\right)-\frac{1}{|\mathcal{H}|}\left(\sum_{C\in\mathcal{H}\setminus\mathcal{R}}\ell_H(f_H^*,C)+\sum_{C\in\mathcal{R}}\ell_M(f_M,C)\right)\\
&=\frac{1}{|\mathcal{H}|}\left(\ell_H(f_H^*,C)-\ell_M(f_M,C)\right)\\
&<0
\end{aligned}$$

so $f_M, \mathcal{R}$ is not an optimal solution to Problem 3. Thus $f_M, \mathcal{R}$ must be a feasible solution to the Problem 1.

Part 2. We now show that if f_M^*, C is an optimal solution to Problem 3, then f_M^*, C is also a solution to Problem 2. We first simplify Problem 3.

$$\begin{aligned}
\text{Problem 3} &\equiv \min_{\mathcal{R}}\min_{f_M}\frac{1}{|\mathcal{H}|}\left(\sum_{C\in\mathcal{H}\setminus\mathcal{R}}\ell_H(f_H^*,C)+\sum_{C\in\mathcal{R}}\ell_M(f_M,C)\right)\\
&\equiv \min_{\mathcal{R}}\frac{1}{|\mathcal{H}|}\sum_{C\in\mathcal{H}\setminus\mathcal{R}}\ell_H(f_H^*,C)+\min_{f_M}\frac{1}{|\mathcal{H}|}\sum_{C\in\mathcal{R}}\ell_M(f_M,C)\\
&\equiv \min_{\mathcal{R}}\frac{1}{|\mathcal{H}|}\left(\sum_{C\in\mathcal{H}\setminus\mathcal{R}}\ell_H(f_H^*,C)+\sum_{C\in\mathcal{R}}\ell_M(f_M^{\mathcal{R}},C)\right)\\
&\equiv \text{Problem 2}.
\end{aligned}$$

This is because

$$f_M^{\mathcal{R}} = \arg\min_{f_M}\frac{1}{|\mathcal{H}|}\sum_{C\in\mathcal{R}}\ell_M(f_M,C),$$

since squared loss is uniquely minimized by the mean.

Thus Problem 3 is equivalent to Problem 2 and we are done. □

B.2 Proof of Proposition 2

Recall Proposition 2.

Proposition 2. *Define a matrix $V \in \mathbb{R}^{|\mathcal{H}|} \times \mathbb{R}^{|\mathcal{M}|}$ with entries $v_{ij} = f^*(\mathbf{x}_{ij})$. The problem of finding an optimal delegate is as follows:*

> VARIANCEASSIGNMENT. Fix a set of rows S of V. For each row $i \in S$, pay a cost proportional to the variance of V across row i, and remove row i from V. Then, for each column j, pay a cost proportional to the variance across column j of the remaining entries. Find a set S^* that minimizes the total cost.

Then for $\mathcal{R} = \{C_i : i \notin S^*\}$, $f_M^{\mathcal{R}}$ will be an optimal delegate.

To formalize this, we must define the variance of a set. For a finite set $S \subset \mathbb{R}$, define the mean μ of S as $\mu(S) := \frac{1}{|S|}\sum_{x\in S} x$. The variance σ^2 of S is $\sigma^2(S) := \frac{1}{|S|}\sum_{x\in S}(x-\mu(x))^2$. When we take $\sigma^2(f^*(\mathbf{x}) : \mathbf{x} \in X)$ for some set of states X, we will simply write $\sigma^2(f^*|X)$.

Finally, if $\mathcal{R}$ is a set of retained human categories, define $S(\mathcal{R})$ to be the set of all states in categories in $\mathcal{R}$, $S(\mathcal{R}) := \bigcup_{C\in\mathcal{R}} C$.

We now formalize and extend Proposition 2 in the following Proposition.

Proposition. *To find an optimal delegate f_M^*, it is sufficient to find a set $\mathcal{R}$ that solves*

$$\min_{\mathcal{R}}\frac{1}{|\mathcal{H}|}\sum_{C\in\mathcal{H}\setminus\mathcal{R}}\sigma^2(f^*|C)+\frac{1}{n}\left(\sum_{K}|K\cap S(\mathcal{R})|\sigma^2\left(f^*|K\cap S(\mathcal{R})\right)\right)$$

and take $f_M^* = f_M^{\mathcal{R}}$. If the human and machine do not share any features, $I_H \cap I_M = \emptyset$, then there is a single $\mathbf{x}_{ij} \in C_i \cap K_j$ for each i, j, and we can define $v_{ij} = f^*(\mathbf{x}_{ij})$ to form a matrix $V \in \mathbb{R}^{|\mathcal{H}|} \times \mathbb{R}^{|\mathcal{M}|}$. The problem of finding an optimal delegate is as follows:

> Fix a set of rows $R \subset [|\mathcal{H}|]$. For each row $i \notin R$, pay a cost $\frac{1}{|\mathcal{H}|}\sigma^2(v_{ij} : j \in [|\mathcal{M}|])$.
>
> For each column, pay a cost $\frac{|R|}{n}\sigma^2(v_{ij} : i \in R)$. Find the set R^* that minimizes the total cost.

Then for $\mathcal{R} = \{C_i : i \in R^*\}$, $f_M^{\mathcal{R}}$ will be an optimal delegate.

Proof. Recall from Proposition 1 that to find an optimal delegate f_M^*, it is sufficient to find $\mathcal{R}$ that solves

$$\arg\min_{\mathcal{R} \subseteq \mathcal{H}} \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \ell_H(f_H^*, C) + \sum_{C \in \mathcal{R}} \ell_M(f_M^{\mathcal{R}}, C) \right). \tag{4}$$

and take $f_M^* = f_M^{\mathcal{R}}$. Thus for the remainder of this proof, we will focus on solving the problem above.

First, note that each category C has size $|C| = \frac{n}{|\mathcal{H}|}$.

Substituting in the expressions for ℓ_H and ℓ_M, we have the objective

$$\begin{aligned}
&\frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{H} \setminus \mathcal{R}} \sum_{\mathbf{x} \in C} \frac{1}{|C|} (f_H^*(C) - f^*(\mathbf{x}))^2 + \sum_{C \in \mathcal{R}} \sum_{\mathbf{x} \in C} \frac{1}{|C|} (f_M^{\mathcal{R}}(K(\mathbf{x})) - f^*(\mathbf{x}))^2 \right) \\
&= \frac{1}{|\mathcal{H}|} \sum_{C \in \mathcal{H} \setminus \mathcal{R}} \sigma^2(f^*|C) + \frac{1}{|\mathcal{H}|} \left(\sum_{C \in \mathcal{R}} \sum_{\mathbf{x} \in C} \frac{|\mathcal{H}|}{n} (f_M^{\mathcal{R}}(K(\mathbf{x})) - f^*(\mathbf{x}))^2 \right) && \text{(by defn of } f_H^*) \\
&= \frac{1}{|\mathcal{H}|} \sum_{C \in \mathcal{H} \setminus \mathcal{R}} \sigma^2(f^*|C) + \frac{1}{n} \left(\sum_{K} \sum_{\mathbf{x} \in K \cap S(\mathcal{R})} (f_M^{\mathcal{R}}(K) - f^*(\mathbf{x}))^2 \right) \\
&= \frac{1}{|\mathcal{H}|} \sum_{C \in \mathcal{H} \setminus \mathcal{R}} \sigma^2(f^*|C) + \frac{1}{n} \left(\sum_{K} |K \cap S(\mathcal{R})| \sum_{\mathbf{x} \in K \cap S(\mathcal{R})} \frac{1}{|K \cap S(\mathcal{R})|} (f_M^{\mathcal{R}}(K) - f^*(\mathbf{x}))^2 \right) \\
&= \frac{1}{|\mathcal{H}|} \sum_{C \in \mathcal{H} \setminus \mathcal{R}} \sigma^2(f^*|C) + \frac{1}{n} \left(\sum_{K} |K \cap S(\mathcal{R})| \sigma^2(f^*|K \cap S(\mathcal{R})) \right) && \text{(by defn of } f_M^{\mathcal{R}})
\end{aligned}$$

If the human and machine features partition the set of features, then $|K \cap C| = 1$ for each human category C and machine category K. Then

$$|K \cap S(\mathcal{R})| \triangleq |K \cap \bigcup_{C \in \mathcal{R}} C| = |\bigcup_{C \in \mathcal{R}} K \cap C| = \sum_{C \in \mathcal{R}} 1 = |\mathcal{R}|.$$

We can thus simplify the previous expression to

$$\frac{1}{|\mathcal{H}|} \sum_{C \in \mathcal{H} \setminus \mathcal{R}} \sigma^2(f^*|C) + \frac{|\mathcal{R}|}{n} \sum_{K} \sigma^2(f^*|K \cap S(\mathcal{R}))$$

Let $R(\mathcal{R}) = \{i : C_i \in \mathcal{R}\}$. Given V as defined above,

$$\sigma^2(f^*|C_i) = \sigma^2(f^*(\mathbf{x}) : \mathbf{x} \in C_i) = \sigma^2(v_{ij} : j \in [|\mathcal{M}|]),$$

and

$$\sigma^2(f^*|K_j \cap S(\mathcal{R})) = \sigma^2(v_{ij} : i \in R(\mathcal{R}))$$

We can write our problem as

$$\min_{\mathcal{R}} \frac{1}{|\mathcal{H}|} \sum_{i \notin \mathcal{R}} \sigma^2(v_{ij} : j \in [|\mathcal{M}|]) + \frac{|\mathcal{R}|}{n} \sum_{j} \sigma^2(v_{ij} | i \in R(\mathcal{R}))$$

Since there is a bijective relationship between R and $\mathcal{R}$, we can simply optimize to find R. □

B.3 Proof of Theorem 3

Recall Theorem 3, which showed that for functions that are separable into functions of the human features and machine features respectively, we may find an optimal delegate in polynomial time in the size of f^*, which is n.

Theorem 3. *Suppose that f^* is separable, that is, $f^*(\mathbf{x}) = u(C(\mathbf{x})) + w(K(\mathbf{x}))$ for some functions u, w. Then we can find an optimal delegate f_M^* in time polynomial in the size of f^*.*

We first provide a brief sketch of the proof. When the human and machine partition the set of all features, we can write $v_{ij} = f^*(\mathbf{x}_{ij}) = u_i + w_j$ for each i, j, for some $u_i, w_j \in \mathbb{R}$. By Proposition 2, we need to find some set R solving

$$\min_R \frac{1}{|\mathcal{H}|} \sum_{i \notin R} \sigma^2(v_{ij} : j \in [|\mathcal{M}|]) + \frac{|R|}{n} \sum_j \sigma^2(v_{ij} | i \in R)$$

In the separable case, we may simplify this to

$$\min_k \left[\left(1 - \frac{k}{|\mathcal{H}|}\right) \cdot \sigma^2(w_j : j \in [|\mathcal{M}|]) + \left(\frac{k}{|\mathcal{H}|}\right) \cdot \min_{R:|R|=k} \sigma^2(u_i : i \in R) \right].$$

The problem is now simply to find the minimum variance subset of size k for each $1 \leq k \leq h$, which we can do efficiently. If the human and machine share features, then we can solve the problem by finding the optimal function by solving a separable sub-problem with independent human and machine features for each of the $s \leq h$ possible values of the shared features.

Proof. Let $h := |\mathcal{H}|$, $m := |\mathcal{M}|$, and define $u_i := u(C_i)$, $w_j := w(K_j)$.

Since we assume that the set of features shared by the human and machine is empty, $n = hm$. Moreover, in this case, we can write $v_{ij} = f^*(\mathbf{x}_{ij}) = u_i + w_j$. We previously did not specify how we indexed the human categories $\mathcal{H}$; we may now index $\mathcal{H}$ so that $u_1 \leq u_2 \leq ... \leq u_h$; performing this indexing has polynomial time complexity $O(h \log h) = O(n \log n)$ since $h = O(n)$ in the worst case.

By Proposition 2, we need to find some set R solving

$$\begin{aligned}
&\min_R \frac{1}{h} \sum_{i \notin R} \sigma^2(v_{ij} : j \in [m]) + \frac{|R|}{n} \sum_j \sigma^2(v_{ij} | i \in R) \\
&= \min_R \frac{1}{h} \sum_{i \notin R} \sigma^2(u_i + w_j : j \in [m]) + \frac{|R|}{n} \sum_j \sigma^2(u_i + w_j | i \in R) \\
&= \min_R \frac{1}{h} \sum_{i \notin R} \sigma^2(w_j : j \in [m]) + \frac{|R|}{n} \sum_j \sigma^2(u_i | i \in R) \\
&= \min_R \frac{h - |R|}{h} \sigma^2(w_j : j \in [m]) + \frac{|R|}{n} m \sigma^2(u_i | i \in R) \\
&= \min_R \left(1 - \frac{|R|}{h}\right) \sigma^2(w_j : j \in [m]) + \frac{|R|}{n} \frac{n}{h} \sigma^2(u_i | i \in R) \\
&= \min_k \min_{R:|R|=k} \left(1 - \frac{k}{h}\right) \sigma^2(w_j : j \in [m]) + \frac{k}{h} \sigma^2(u_i | i \in R) \\
&= \min_k \left(1 - \frac{k}{h}\right) \sigma^2(w_j : j \in [m]) + \frac{k}{h} \min_{R:|R|=k} \sigma^2(u_i | i \in R)
\end{aligned}$$

where the last step was suggested in [29]. Let

$$R_k \in \operatorname*{arg\,min}_{R \subseteq [h]:|R|=k} \sigma^2(u_i | i \in R).$$

Then our objective is

$$\min_k \left[\left(1 - \frac{k}{h}\right) \cdot \sigma^2(w_j : j \in [m]) + \left(\frac{k}{h}\right) \cdot \sigma^2(u_i | i \in R_k) \right].$$

We may now outline an algorithm to compute the optimal retained rows R^*. First, we compute R_k for each k. We then iterate over each $1 \leq k \leq h$ and find the k^* that minimizes $\left(1 - \frac{k}{h}\right) \sigma^2(w_j : j \in [m]) + \left(\frac{k}{h}\right) \sigma^2(u_i : i \in R_k)$. Then R_{k^*} minimizes the objective. Take $\mathcal{R} = \{C_i : i \in R_{k^*}\}$. By Proposition 2, $f_M^{\mathcal{R}}$ is an optimal machine design.

Given $\{R_k\}_{k=1}^h$, finding k^* simply requires computing the objective for each $1 \leq k \leq h$, which has total time complexity $O(h)$. It therefore only remains to find a polynomial-time algorithm to compute R_k. This problem is equivalent to finding the minimum variance subset of u of size k, and setting R_k to be the indices corresponding to that subset. This can be done by observing that the minimum variance subset must be *contiguous*. This was previously conjectured [29]; we prove this formally in Lemma 1 using a proof technique similar that suggested by [29]. We can then compute the variance of each of the $h - k + 1$ contiguous subsets of $\{u_i\}$ in time $O(h)$, for a total time of $O(h^2)$. □

Lemma 1. *For each k, there is some $1 \leq t \leq h - k + 1$ such that*

$$\{t, t+1, ..., t+k-1\} \in \underset{R \subseteq [h]:|R|=k}{\arg\min} \sigma^2(u_i : i \in R).$$

Proof of Lemma 1. Fix k, and suppose that there is no contiguous minimum variance subset of size k. Let $R_k \in \arg\min_{R \subseteq [h]:|R|=k} \sigma^2(u_i : i \in R)$. Let $i = \min R_k, i' = \max R_k$. We may assume without loss of generality that there is no $u_j = u_i$ for $j > i$ and $j \notin R_k$; otherwise replace R_k with $R_k \cup \{j\} \setminus \{i\}$, which will also be a minimum variance subset. Similarly we may assume that there is no $u_j = u_{i'}$ for $j < i'$ and $j \notin R_k$.

Denote $\bar{u}^S = \mu(u_s : s \in S), \sigma^2(S) = \sigma^2(u_s : s \in S)$.

By assumption, R_k is not contiguous, that is, there is some $i < j < i'$ such that $j \notin R_k$ and $u_i < u_j < u_{i'}$. Suppose that $u_j \leq \bar{u}^{R_k}$. Let $R_k^0 := R_k \setminus \{i\}$. Since $u_i \leq u_s$ for all $s \in I_k$, $\bar{u}^{R_k^0} \geq \bar{u}^{R_k}$.

Now, let $R_k' := R_k^0 \cup j$; essentially R_k' is the result of replacing C_i with C_j in R_k. By altering Result 1 of [30] for the true variance rather than the sample variance, we see that

$$k\sigma^2(R_k) = 1 + (k-1)\sigma^2(R_k^0) + \frac{k-1}{k}\left(u_i - \bar{u}^{R_k^0}\right)^2$$

and

$$k\sigma^2(R_k') = 1 + (k-1)\sigma^2(R_k^0) + \frac{k-1}{k}\left(u_j - \bar{u}^{R_k^0}\right)^2$$

Thus

$$\begin{aligned}
\sigma^2(R_k') - \sigma^2(R_k) &= \frac{k-1}{k^2}\left(\left(u_j - \bar{u}^{R_k^0}\right)^2 - \left(u_i - \bar{u}^{R_k^0}\right)^2\right) \\
&= \frac{k-1}{k^2}\left(\left(u_j - \bar{u}^{R_k^0}\right) + \left(u_i - \bar{u}^{R_k^0}\right)\right)\left(\left(u_j - \bar{u}^{R_k^0}\right) - \left(u_i - \bar{u}^{R_k^0}\right)\right) \\
&= \frac{k-1}{k^2}\left(u_j - \bar{u}^{R_k^0} + u_i - \bar{u}^{R_k^0}\right)(u_j - u_i)
\end{aligned}$$

Since $u_j < u_i$ and $u_i, u_j \leq \bar{u}^{R_k} \leq \bar{u}^{R_k^0}$, $\sigma^2(R_k') - \sigma^2(R_k) < 0$. Thus R_k is not the minimum variance subset, and we have a contradiction.

If $u_j > \bar{u}^{R_k}$, we can let $R_k^0 := R_k \setminus \{C_{i'}\}$ and $R_k' := R_k^0 \cup \{C_j\}$, which by a symmetric argument again yields $\sigma^2(R_k') - \sigma^2(R_k) < 0$. □

B.4 Proof of Theorem 4

Recall Theorem 4.

Theorem 4. *Suppose that $|I_H| = O(1)$ or $|I_M| = O(1)$. Then we may find an optimal delegate in time polynomial in the size of f^*.*

We prove each case individually. Recall that size$(f^*) = n$, the number of states, and again let $h := |\mathcal{H}|, m := |\mathcal{M}|$.

Lemma 2. *If $|I_H| = O(1)$, we may find an optimal delegate in time polynomial in the size of f^*.*

Proof. By Proposition 1, it is sufficient to find $\mathcal{R} \subseteq \mathcal{H}$ that minimizes $\ell(f_H^*, f_M^{\mathcal{R}})$. If $|I_H| = O(1)$, then $|\mathcal{H}| = 2^{|I_H|} = O(1)$, and there are $2^{O(1)} = O(1)$ subsets of $\mathcal{H}$. Moreover, we may compute $f_M^{\mathcal{R}}$ and then $\ell(f_H^*, f_M^{\mathcal{R}})$ in $O(n)$ for each $\mathcal{R}$. Thus, it is sufficient to take a brute force approach and check the loss of $f_M^{\mathcal{R}}$ for each $\mathcal{R} \subseteq \mathcal{H}$ in time $O(n)$. □

Lemma 3. *If $|I_M| = O(1)$, we may find an optimal delegate in time polynomial in the size of f^*.*

Proof. Since we assume that I_H and I_M partition the set of all features, there is a single state $\mathbf{x}_{ij}$ in $C_i \cap K_j$ for each human category C_i and machine category K_j.

For any machine function f_M, we may define the vector $\mathbf{y} \in \mathbb{R}^m$ by setting $\mathbf{y}_j = f_M(K_j)$. Define $c_{ij} = f^*(\mathbf{x}_{ij})$, and $r_i^2 = \sum_j (f_H^*(C_i) - f^*(x_{ij}))^2$.

We know that a human with action function f_H^* will delegate to a machine with action function f_M in category C_i if and only if

$$\sum_{j=1}^{m} \frac{1}{m}(f_M(K_j) - f^*(\mathbf{x}_{ij}))^2 < \sum_{j=1}^{m} \frac{1}{m}(f_H^*(C_i) - f^*(\mathbf{x}_{ij}))^2,$$

assuming that the human breaks ties by not delegating. We can rewrite this condition as

$$\sum_{j=1}^{m} (\mathbf{y}_j - c_{ij})^2 < r_i^2,$$

which is the equation for the interior of an high-dimensional sphere where $\mathbf{y} \in \mathbb{R}^m$. This means that for any machine f_M, f_M is in the region formed by the intersection of the spheres corresponding to the categories where f_M is adopted, $\mathcal{D}(f_H^*, f_M)$.

If $f^*(\mathbf{x})$ is rational for all $\mathbf{x}$ – which is a reasonable assumption if we hope top optimize on a computer with floating point precision – then the sphere $g_i(\mathbf{y}) = \sum_{j=1}^{m} (\mathbf{y}_j - c_{ij}^2) - r_i^2$ is a polynomial with maximum degree $2m$ in $\mathbb{R}^m$, and $g_i(y) = 0$ is the sphere corresponding to category C_i. The regions of intersection of these h spheres are known as "arrangements" in computational geometry. There are only $O(h^m)$ such regions, and these regions may be found in time $O(h^m)$ [31].

Thus the first step of this algorithm will be to find these regions, which takes $O(h^m) = O(n^m) = O(\text{poly}(n))$ since $h = O(n)$.

In Proposition 1, we showed that to find an optimal delegate, it is sufficient to find a set of categories $\mathcal{R}^*$ such that $f_M^{\mathcal{R}}$ minimizes the team loss, and $\mathcal{R}^*$ satisfies $\mathcal{R}^* = \mathcal{D}(f_H^*, f_M^*)$ for some optimal delegate f^*. This means that $C_i \in \mathcal{R}^*$ if and only if f_M^* is in sphere i. This in turn implies that in searching over different sets of retained categories $\mathcal{R}$, we can consider only subsets of categories whose corresponding set of spheres has a non-empty intersection.

Now, instead of 2^h possible options for $\mathcal{R}$, we are only searching over $O(h^m)$ different subsets $\mathcal{R}$. Since $m = O(1)$ and $h = O(n)$, $O(h^m)$ is polynomial in n. Moreover, since we can also compute the loss $f_M^{\mathcal{R}}$ of a given $\mathcal{R}$ in $O(n)$, we may find the optimal delegate in time polynomial in n. □

B.5 Proof of Theorem 5

We now prove that finding an optimal delegate is NP-hard in general. Recall from Proposition 2 that to find an optimal delegate it is necessary to solve the problem VARIANCEASSIGNMENT.

Proposition 2. *Define a matrix $V \in \mathbb{R}^{|\mathcal{H}|} \times \mathbb{R}^{|\mathcal{M}|}$ with entries $v_{ij} = f^*(\mathbf{x}_{ij})$. The problem of finding an optimal delegate is as follows:*

> VARIANCEASSIGNMENT. Fix a set of rows S of V. For each row $i \in S$, pay a cost proportional to the variance of V across row i, and remove row i from V. Then, for each column j, pay a cost proportional to the variance across column j of the remaining entries. Find a set S^* that minimizes the total cost.

Then for $\mathcal{R} = \{C_i : i \notin S^*\}$, $f_M^{\mathcal{R}}$ will be an optimal delegate.

We will show that VARIANCEASSIGNMENT is NP-hard.

First, consider the problem MAXREGULARCLIQUE.

> MAXREGULARCLIQUE. Let $G = (V, E)$ be a regular graph. A clique is a set of nodes $S \subseteq V$ such that for each pair of nodes $u, v \in S$, $(u, v) \in E$. Find a clique with maximum size $|S|$.

[32] defines the problem REGULARCLIQUE, which determines whether a regular graph G contains a clique of size k. They show that it is NP-hard in their Theorem 3. We can solve REGULARCLIQUE by solving MAXREGULARCLIQUE and checking whether the solution has size $\geq k$; thus MAXREGULARCLIQUE is also NP-hard.

We now define the intermediate problem of densest subgraph discovery in the presence of possibly negative weights and a regularity condition on each node.

> NEGREGULARDSD. Let $G = (V, E, w)$ be an undirected graph with weighted edges $w : E \to \mathbb{R}$. Suppose that for each node v,
>
> $$\sum_{(u,v)\in E} |w(u, v)| = 1.$$
>
> For a subset of nodes $S \subseteq V$, let $E(S)$ be the edges in the induced subgraph, and define
>
> $$w(S) = \sum_{(u,v)\in E(S)} w(u, v).$$
>
> Find the subset of nodes S that maximizes the density of the induced subgraph
>
> $$d(S) = \frac{w(S)}{|S|},$$
>
> where $d(S) = 0$ when $S = \emptyset$.

Without the regularity condition that the absolute sum of a node's edge weights is equal to 1, this is the NP-hard problem NEGDSD introduced in [33]. It is also folk knowledge that NEGDSD may be proved via a reduction from MAXCLIQUE. We now show that NEGREGULARDSD is NP-hard via a reduction from MAXREGULARCLIQUE.

Proof. Let the d-regular graph $G = (V, E)$ be an instance of MAXREGULARCLIQUE, we construct a complete graph $G' = (V', E', w)$ where $V' = V$ and E' is the set of all pairs of nodes. Let

$$w(u, v) = \frac{1}{1 + (n - d)} \cdot \begin{cases} \frac{1}{d}, & (u, v) \in E, \\ -1, & (u, v) \notin E \end{cases}$$

Constructing G' can be completed in polynomial time $O(|V|^2)$.

Now

$$\sum_{(u,v)\in E'} |w(u, v)| = d \cdot \frac{1}{d} \cdot \frac{1}{1 + (n - d)} + (n - d) \cdot 1 \cdot \frac{1}{1 + n - d} = 1.$$

First note that the solution S to NEGREGULARDSD will always have non-negative density, since we could always pick the empty set.

Now, suppose there is a solution S to NEGREGULARDSD that is not a clique in G. Then there is some pair $u, v \in S$ such that $(u, v) \notin E$. Let $E(N)$ be the edges in G' induced by a set of nodes N.

Then,

$$\begin{aligned}
d(S) &= \frac{\sum_{(s,t)\in E(S)}}{|S|} \\
&= \frac{\sum_{(s,t)\in E(S),t\neq v} w(s,t) + \sum_{(s,v)\in E(S),s\neq u} w(s,v) + w(u,v)}{|S|} \\
&\leq \frac{\sum_{(s,t)\in E(S),t\neq v} w(s,t) + \sum_{(s,v)\in E(S)s\neq u} w(s,v) - \frac{1}{1+n-d}}{|S|-1} \\
&\leq \frac{\sum_{(s,t)\in E(S)t\neq v} w(s,t) + d\cdot\frac{1}{d}\cdot\frac{1}{1+(n-d)} - \frac{1}{1+n-d}}{|S|-1} \\
&= \frac{\sum_{(s,t)\in E(S\setminus\{v\})}}{|S|-1} \\
&= d(S\setminus\{v\}),
\end{aligned}$$

so $d(S)$ cannot have been a solution of NEGREGULARDSD. Thus NEGREGULARDSD will produce a solution which is a clique in G. For a subset $S \subset V$ that is a clique in G, the density is

$$\begin{aligned}
d(S) &= \frac{\sum_{(u,v)\in E(S)} w(u,v)}{|S|} \\
&= \frac{\binom{|S|}{2}\frac{1}{d}\cdot\frac{1}{1+(n-d)}}{|S|} \\
&= \frac{1}{d}\cdot\frac{1}{1+(n-d)}\left(\frac{|S|(|S|-1)}{2}\right)\frac{1}{|S|} \\
&= \frac{1}{d}\cdot\frac{1}{1+(n-d)}\cdot\frac{|S|-1}{2} \\
&\propto |S|-1
\end{aligned}$$

Thus NEGREGULARDSD will select the clique of maximum size. □

Finally, reduce NEGREGULARDSD to VARIANCEASSIGNMENT.

Proof. Let $G = (V, E, w)$ be an instance of NEGREGULARDSD. Construct an instance A of VARIANCEASSIGNMENT as follows.

First, create a matrix A^0: for each node $v_i \in V$, create a row i; for each edge $e_j = (i,k)$ create a column j.

For each edge $e_j = (v_i, v_k)$ if $w(v_i, v_k) > 0$, let $a^0_{ij} = a^0_{ik} = \sqrt{|w(i,k)|/2}$. If $w(v_i, v_k) \leq 0$, let $a^0_{ij} = -a^0_{ik} = \sqrt{|w(v_i, v_k)|/2}$. Set all other entries to zero.

Now, create a matrix A as follows: for each column a_j of A^0, add both a_j and $-a_j$ to A.

Constructing A takes time $O(2|E||V|) = O(|V|^3)$.

Let $m = 2|E|$ and $h = |V|$, then $A \in \mathbb{R}^h \times \mathbb{R}^m$ and $n = hm$.

Now for each row i,

$$\sum_j a_{ij} = 0$$

and

$$\|a_i\|_2^2 = \sum_j a_{ij}^2 = 2\sum_{k:(v_i,v_k)\in E} |w(v_i, v_k)|/2 = 1.$$

Let R be a subset of rows. Then the objective of VARIANCEASSIGNMENT is to minimize

$$\frac{1}{h}\sum_{i\notin R}\sigma^2(a_{ij} : j \in [m]) + \frac{R}{n}\sum_j \sigma^2(a_{ij}|i \in R).$$

We may expand this to

$$\begin{aligned}
&\frac{1}{h}\sum_{i\notin R}\mu(a_{ij}^2 : j\in[m]) - \mu(a_{ij} : j\in[m])^2 + \frac{|R|}{n}\sum_j \mu(a_{ij}^2|i\in R) - \mu(a_{ij}|i\in R)^2 \\
&= \frac{1}{h}\sum_{i\notin R}\mu(a_{ij}^2 : j\in[m]) + \frac{|R|}{n}\sum_j \mu(a_{ij}^2|i\in R) \\
&\quad - \left(\frac{1}{h}\sum_{i\notin R}\mu(a_{ij} : j\in[m])^2 + \frac{|R|}{n}\sum_j \mu(a_{ij}|i\in R)^2\right)
\end{aligned}$$

The first two terms can be simplified as

$$\begin{aligned}
&\frac{1}{h}\sum_{i\notin R}\mu(a_{ij}^2 : j\in[m]) + \frac{|R|}{n}\sum_j \mu(a_{ij}^2|i\in R) \\
&= \frac{1}{h}\sum_{i\notin R}\frac{1}{m}\sum_j a_{ij}^2 + \frac{|R|}{n}\sum_j \frac{1}{|R|}\sum_{i\in R} a_{ij}^2 \\
&= \frac{1}{n}\sum_{i\notin R}\sum_j a_{ij}^2 + \frac{1}{n}\sum_{i\in R}\sum_j a_{ij}^2 \\
&= \frac{1}{n}\sum_{i,j} a_{ij}^2
\end{aligned}$$

This term is constant in R, so the problem of VARIANCEASSIGNMENT is merely the problem of minimizing the second two terms, or *maximizing*

$$\frac{1}{h}\sum_{i\notin R}\mu(a_{ij} : j\in[m])^2 + \frac{|R|}{n}\sum_j \mu(a_{ij}|i\in R)^2.$$

We know that in this instance,

$$\frac{1}{h}\sum_{i\notin R}\mu(a_{ij} : j\in[m])^2 = \frac{1}{h}\sum_{i\notin R}\left(\frac{1}{m}\sum_j a_{ij}\right)^2 = 0,$$

so – ignoring the factor of $1/n$ – we are really maximizing

$$\begin{aligned}
|R|\sum_j \mu(a_{ij}|i\in R)^2 &= |R|\sum_j\left(\frac{1}{|R|}\sum_{i\in R} a_{ij}\right)^2 \\
&= \frac{1}{|R|}\sum_j\left(\sum_{i\in R} a_{ij}\right)^2 \\
&= \frac{1}{|R|}\sum_j\sum_{i\in R} a_{ij}^2 + 2\sum_{i<k} a_{ij}a_{kj} \\
&= \frac{1}{|R|}\sum_{i\in R}\sum_j a_{ij}^2 + 2\sum_{i<k}\sum_j a_{ij}a_{kj} \\
&= \frac{1}{|R|}\sum_{i\in R}\|a_i\|^2 + 2\frac{1}{|R|}\sum_{i<k}\sum_j a_{ij}a_{kj} \\
&= 1 + 2\frac{1}{|R|}\sum_{i<k:i,k\in R}\sum_j a_{ij}a_{kj}
\end{aligned}$$

so in solving VARIANCEASSIGNMENT we are maximizing $\sum_{i<k}\sum_j a_{ij}a_{kj}$. Recalling the construction of A, $a_{ij}a_{kj} \neq 0$ if and only if $e_j = (v_i, v_k)$. If this is the case, $a_{ij}a_{kj} = w(v_i, v_k)$. Let $S(R) = \{v_i : i \in R\}$ Thus the objective simplifies to

$$\frac{\sum_{(v_i,v_k)\in E(S(R))} w(v_i, v_k)}{|R|} = d(S(R))$$

and in minimizing the VARIANCEASSIGNMENT objective we are maximizing $d(S)$. □

Thus VARIANCEASSIGNMENT is NP-hard.

C Additional features configurations

In the main text, we assumed that the human features I_H and machine features I_M formed a partition of set of all features $[d]$, but our theoretical results hold more generally.

First, suppose that in a delegation setting there is some set of features that are observed by neither the human or the machine. Then the problem of finding an optimal delegate is equivalent to one in which the human and machine together have access to all features. We show this below formally.

Proposition 6. *Let $\mathbf{x}_{H\cup M}$ denote $\mathbf{x}$ restricted to the features in $I_H \cup I_M$. Given a ground truth optimal action function f^*, for $\mathbf{x} \in C_i \cap K_j$ define*

$$\bar{f}(\mathbf{x}_{H\cup M}) = \frac{1}{|C_i \cap K_j|} \sum_{\mathbf{z}\in C_i\cap K_j} f^*(\mathbf{z}).$$

Then a machine function f_M^* is an optimal delegate for ground truth function f^* if and only if f_M^* is an optimal delegate for ground truth function $\bar{f}$.

Proof. A human or machine agent A with action function f_A will have expected loss in human category C in a delegation setting with ground truth optimal action function f^* of

$$\begin{aligned}
&\frac{1}{|C|}\sum_K \sum_{\mathbf{x}\in C\cap K} (f_A(\mathbf{x}) - f^*(\mathbf{x}))^2 \\
&= \frac{1}{|C|}\sum_K \sum_{\mathbf{x}\in C\cap K} (f_A(\mathbf{x}) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - f^*(\mathbf{x}))^2 \\
&= \frac{1}{|C|}\sum_K \sum_{\mathbf{x}\in C\cap K} (f_A(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \\
&\quad - \frac{1}{|C|}\sum_K \sum_{\mathbf{x}\in C\cap K} 2(f_A(\mathbf{x}) - \bar{f}(\mathbf{x}))(\bar{f}(\mathbf{x}) - f^*(\mathbf{x})) \\
&\quad + \frac{1}{|C|}\sum_K \sum_{\mathbf{x}\in C\cap K} (\bar{f}(\mathbf{x}) - f^*(\mathbf{x}))^2 \\
&= (1) - (2) + (3)
\end{aligned}$$

Term (2) can be simplified, since f_A and $\bar{f}$ are constant in $C \cap K$:

$$\begin{aligned}
&\frac{1}{|C|}\sum_K \sum_{\mathbf{x}\in C\cap K} 2(f_A(\mathbf{x}) - \bar{f}(\mathbf{x}))(\bar{f}(\mathbf{x}) - f^*(\mathbf{x})) \\
&= 2\frac{1}{|C|}\sum_K (f_A(\mathbf{x}) - \bar{f}(\mathbf{x})) \sum_{\mathbf{x}\in C\cap K} (\bar{f}(\mathbf{x}) - f^*(\mathbf{x})) \\
&= 2\frac{1}{|C|}\sum_K (f_A(\mathbf{x}) - \bar{f}(\mathbf{x})) \cdot 0 \qquad \text{(definition of } \bar{f}\text{)} \\
&= 0
\end{aligned}$$

Term (1) is the expected loss in category C of using function f_A in category C when the ground truth action function is $\bar{f}$, and term (3) is a constant independent of f_A. Thus the team loss with ground

truth function f^* will be different from the team loss with ground truth function $\bar{f}$ by a constant factor of $\sum_C \frac{1}{|C|} \sum_K \sum_{\mathbf{x} \in C \cap K} (\bar{f}(\mathbf{x}) - f^*(\mathbf{x}))^2$, and the minimization problems have the same set of solutions. □

We can therefore assume without loss of generality that the human and machine together observe all features.

Now, we consider the case when the human and the machine may share some features $I_S = I_H \cap I_M$.

The proof of Proposition 1 in Section B makes no assumptions on the relation between the human and machine features, and thus generalizes entirely.

The proof of Proposition 2 in Section B shows that when the human and machine share features, the insight that the relative *variances* is the crucial factor in determining the optimal machine design remains true.

Theorem 5 generalizes vacuously: since the problem is hard in the restricted setting, the problem must be hard in the more general setting.

In the definition of separable functions, we made no assumptions on the overlap between human and machine features, and indeed Theorem 3 generalizes entirely. We prove this below.

Proof that Theorem 3 generalizes. Let $\mathbf{x}_S$ be the values of the features in I_S; let $\mathbf{x}_{H \setminus S}$ be values of the non-shared human features, $\mathbf{x}_{M \setminus S}$ be the values of the non-shared machine features, and $\mathbf{x}_{H \cup M \setminus S}$ be the values of all non-shared features.

There are $s \leq h \leq n$ unique settings of $\mathbf{x}_S$. Label these as $\mathbf{x}_S^{(t)}$ for $1 \leq t \leq s$.

Recall that if f^* is separable, we may write it as

$$f^*(\mathbf{x}) = u(C(\mathbf{x})) + w(K(\mathbf{x}))$$

for some functions u, w; equivalently we may write it as

$$f^*(\mathbf{x}) = u(\mathbf{x}_H) + w(\mathbf{x}_M)$$

for the same functions u, w, since each $\mathbf{x}_H$ or $\mathbf{x}_M$ corresponds to a unique human or machine category respectively.

Since the features I_S are shared, the machine may observe $\mathbf{x}_S^{(t)}$ when selecting an action. We may thus define $f_M(\mathbf{z}_M) = f_{M,t}(\mathbf{z}_{M \setminus S})$ if $\mathbf{z}_S = \mathbf{x}_S^{(t)}$ and select functions $f_{M,\mathbf{x}_S}$ independently.

For $\mathbf{z}$ with $\mathbf{z}_S = \mathbf{x}_S^{(t)}$,

$$f^*(\mathbf{z}) = f_t^*(\mathbf{z}_{H \cup M \setminus S}) = u_t(\mathbf{z}_{H \setminus S}) + w_t(\mathbf{z}_{M \setminus S})$$

where $u_t(\mathbf{z}_{H \setminus S}) := u(\mathbf{x}_S^{(t)}, \mathbf{z}_{H \setminus S}) = u(\mathbf{z}_H)$, $w_t(\mathbf{z}_{M \setminus S}) := w(\mathbf{x}_S^{(t)}, \mathbf{z}_{M \setminus S}) = w(\mathbf{z}_M)$.

Finally, for $\mathbf{z}_S = \mathbf{x}_S^{(t)}$, let $f_{H,t}(\mathbf{z}_{H \setminus S}) = f_H^*(\mathbf{x}_S^{(t)}, \mathbf{z}_{H \setminus S}) = f_H^*(\mathbf{z}_H)$. If the human can only observe $\mathbf{z}_{H \setminus S}$ but $\mathbf{x}_S^{(t)}$ is fixed, then $f_{H,t}^*(\mathbf{z}_{H \setminus S}) = f_{H,t}(\mathbf{z}_{H \setminus S})$ will be the optimal human action in state $\mathbf{z}$, because the human is still simply choosing the optimal action in category $C(\mathbf{z})$.

For each t, find the optimal delegate $f_{M,t}^*$ in the delegation setting where the human has access to $I_H \setminus I_S$, the machine has access to $I_M \setminus I_S$, and the ground truth optimal action is the separable function f_t^*, which can be done in polynomial time. Since there are $O(n)$ such values of t, we can find f_M^* in polynomial time as well. □

Finally, Theorem 4 also generalizes: if $|I_H \setminus I_M| = O(1)$ or $|I_M \setminus I_H| = O(1)$, an optimal delegate can be found in polynomial time. This can be shown through the same argument as for Theorem 3. For each setting of the shared features, we have a subproblem which has human features I'_H and machine features I'_M, where $|I'_H| = |I_H \setminus I_M| = O(1)$ and $|I'_M| = |I_M \setminus I_H| = O(1)$. We can then apply the algorithm from Theorem 4 to find the optimal machine function in each of the $O(n)$ subproblems.