Tractable Agreement Protocols

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth

Abstract

We give an efficient reduction through which any machine learning algorithm can be converted into an interactive protocol that can interact with another party (such as a human) to reach agreement on predictions and improve accuracy. The requirements on each party are calibration conditions which are computationally and statistically tractable relaxations of Bayesian rationality — that are sensible even in prior free settings — and hence are a substantial generalization of Aumann's classic "agreement theorem" Aumann [1976]. In the interactive protocol, the machine learning model first produces a prediction. Then, the human responds to the model's prediction by either conveying agreement, or else providing feedback of some sort. The model then updates its state and provides a new prediction, and the human in turn may update their beliefs. The process continues until the model and the human reach agreement.

The first setting we study generalizes past work on Aumann's Agreement Theorem, in which the parties aim to agree on a one-dimensional expectation. At each round, each party simply communicates an estimate of their current prediction for the expectation. In this setting we recover the quantitative convergence theorem of Aaronson Aaronson [2005] (but under our much weaker assumptions). We then move on to the case in which the parties maintain beliefs about a distribution over d outcomes and consider two feedback mechanisms. The first simply corresponds to a vector-valued estimate of the agents' current prediction. The second takes a decision theoretic perspective: if the human needs to take some downstream action from a finite set, and has an arbitrary utility function of their action and the outcome, then we show that the parties can communicate and reach agreement about the correct downstream action to take by simply communicating at each round the action that they believe to be utility maximizing. The number of rounds until agreement remains independent of d in this case. We can also generalize our protocols to more than 2 parties, with computational complexity that degrades only linearly with the number of parties. Our protocols are based on simple, efficiently maintainable conditions and result in predictions that are more accurate than any single party's alone.

1 Introduction

Consider a machine learning model designed to help doctors make clinical decisions. This predictive model is trained on a much larger dataset of patients than the doctor's experience can draw on. However, it is also necessarily trained on different, and perhaps less rich data. For example, the doctor's observations often include qualitative information, such as their patients' reaction to touch, which are hard to encode as input to the model. As a result, even if the model is more accurate on average than the doctor, there will still be situations in which the doctor ought not to follow the model's recommendation and proceed to act in accordance with their own predictions. In such a situation, rather than forcing the doctor to *either* use the model *or* choose to ignore it, we would prefer an interface through which the doctor and model can interact to update their beliefs and reach agreement on a prediction that is guaranteed to be more accurate than either of their initial

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

predictions. When designing such an interface, we would like to be able to prove convergence and utility guarantees under minimal, computationally and statistically tractable assumptions on the interaction between the model and the doctor, for at least two reasons:

1. We need to actually implement the model's interaction with the protocol. So, we want to be able to start with an arbitrary black-box model (e.g. a trained neural network) and convert it efficiently (in terms of both computation and data requirements) into an interactive system
2. We want to make minimal assumptions on how the human will interact with the model. These assumptions should be significantly weaker than "perfect rationality" — meaning that they should be satisfied by informed Bayesian reasoners — but our results should not hinge on making a computationally implausible assumption. Rather we should make assumptions that can be satisfied in a computationally and statistically tractable way. The weaker our assumptions are, the more likely they are to be satisfied.

In this paper, we show how to efficiently convert an arbitrary predictive model into an interactive protocol that can be used to interact with another party in a way that quickly leads to agreement while improving accuracy under tractable assumptions. We give results across a variety of feedback models and extend our results to multiple parties. Our results rely on the theory of calibration, which has a long intellectual history [Dawid, 1982, 1985, Foster and Vohra, 1998, 1999, Hébert-Johnson et al., 2018, Błasiok et al., 2023], and naturally live in a sequential adversarial setting that involves repeated interaction across many predictions without distributional assumptions. Moreover, if the instances *are* drawn from a prior distribution and the two parties are informed Bayesians, then we are able to give an "online-to-one-shot" reduction that translates our theorems to high probability guarantees for interactions on a single instance drawn from this prior. We show that all of our calibration requirements are satisfied by Bayesians updating on a correct prior, implying that our approach generalizes past works in the well-studied one-shot "Aumann Agreement Theorem" setting [Aumann, 1976, Geanakoplos and Polemarchakis, 1982, Aaronson, 2005, Kong and Schoenebeck, 2023, Frongillo et al., 2023]. In particular, we generalize the kind of instance-independent convergence bounds proven by Aaronson [2005] in the 1-dimensional real valued setting to d dimensions, both when the feedback is vector valued, and when it is only in the form of a "best response action". In the latter case we get convergence at a rate that is not only independent of the complexity of the underlying prior, but also independent of the dimension d.

1.1 Our Model and Results

Over a series of *days* t, examples arrive. Each example has a true label $y^t \in \mathbb{R}^d$ which is initially unobserved and that at least two parties want to predict or otherwise act on (we focus on the two party case for this informal description). We call one of the parties the *human*, and the other the *model*. Before making their initial predictions, the model sees features x_m^t relevant to the example, and the human sees a potentially different set of features x_h^t (potentially over a different feature space). Based on these features, the model and the human engage in a *conversation* over a series of *rounds* $k = 1, \ldots, L$. The human and model alternate speaking, with the model speaking in odd rounds and the human speaking in even rounds. In each odd round k, the model produces a prediction $p_m^{t,k}$ (as a function of all prior history of both conversation rounds that day as well as previous days), which is observed by the human, who in turn produces a prediction $p_h^{t,k+1}$ in round $k+1$, which may also be a function of all previously observed history. The conversation continues until at some round k, the pair of predictions $(p_m^{t,k-1}, p_h^{t,k})$ (if k is even) or $(p_h^{t,k-1}, p_m^{t,k})$ (if k is odd) satisfy an agreement condition, at which point both the human and the model observe the label, and time proceeds to the next day. We give conditions—all of which are computationally and statistically tractable relaxations of full Bayesian rationality—under which the conversation is guaranteed to quickly lead to agreement on predictions that are more accurate than the initial predictions.

The Canonical Setting. In the simplest setting that we study, the labels $y^t \in [0, 1]$ are one dimensional, and the predictions $p_m^{t,k}, p_h^{t,k} \in [0, 1]$ are also one-dimensional numeric values, and intended to convey a numeric estimate of y^t or its expectation. We measure the accuracy of predictions using squared error. For example, the squared error of the human's initial (round 2) predictions over days is $\sum_{t=1}^{T} (p_h^{t,2} - y^t)^2$. In this case, we say that the human and the machine are in ϵ-agreement at some round k if $|p_m^{t,k} - p_h^{t,k-1}| \leq \epsilon$ (odd k) or $|p_h^{t,k} - p_m^{t,k-1}| \leq \epsilon$ (even k). We define a calibration condition that we call "conversation calibration". Informally speaking, conversation calibration for the model requires that for each round k, if we consider the subsequence of days on which conversation extended to round k, denoted by $T^{\geq k}$, the predictions made at round k $\{p_m^{t,k}\}_{t \in T^{\geq k}}$ are calibrated to the outcome subsequence on those days $\{y^t\}_{t \in T^{\geq k}}$ not just marginally, but *conditionally* on the value of the prediction made by the human in the previous round ($k-1$).

Definition 1.1 (Informal, see Definition 3.22)**.** *We say that the model satisfies conversation calibration with respect to the human if for all odd rounds k and $v, v' \in [0,1]$:*

$$\sum_{t \in T^{\geq k}} \mathbb{1}[p_m^{t,k} = v] \cdot \mathbb{1}[p_h^{t,k-1} = v'](p_m^{t,k} - y^t) = 0$$

Conversation calibration for the human is a symmetric condition for even rounds k.

Importantly, conversation calibration does *not* require that the predictions be unbiased conditional on the whole conversation so far (which a correctly specified Bayesian would satisfy), but only conditional on the current prediction of the model, and the most recent prediction of its human interlocutor. This makes the condition computationally and statistically tractable to enforce using standard algorithms for online calibration (e.g. Foster and Vohra [1998], Gupta et al. [2022]). In fact, in our use-case it turns out to be sufficient to measure calibration error using "distance to calibration" Błasiok et al. [2023], a recently defined relaxation of traditional calibration measures. This is useful because there are extremely simple algorithms that can make predictions with "distance to calibration" diminishing at much better rates than are possible for standard calibration metrics Qiao and Zheng [2024], Arunachaleswaran et al. [2025]. When we construct conversation algorithms from static models, we make use of the simple efficient algorithm of Arunachaleswaran et al. [2025], which can be used to bound distance to conversation calibration, as conversation calibration requires conditioning only on disjoint events.

Theorem 1.2 (Informal, see Theorem 4.8)**.** *There is a computationally efficient reduction that takes as input an arbitrary model M mapping features to predictions, and outputs an algorithm that can engage in a conversation protocol. The algorithm uses the predictions of model M at the first round, and is guaranteed to satisfy (approximate) conversation calibration against any agent that it converses with.*

We show that if both parties are conversation calibrated, then on a large fraction of days, the human and the model agree very quickly.

Theorem 1.3 (Informal, see Theorem 4.1)**.** *If the human and model are conversation-calibrated, then for any $\epsilon, \delta \in (0,1]$ and large enough T, on a $1-\delta$ fraction of days, they reach ϵ-agreement after at most $K = \frac{1}{\epsilon^2 \delta}$ rounds of conversation. Furthermore, for this $1-\delta$ fraction of days, if they reach agreement in round i, their final predictions have a lower squared error than the base predictions of either the human or the model, by a term that scales as $\frac{i}{\delta \epsilon^2}$) (so longer conversations directly lead to correspondingly more accurate predictions).*

Paired with our algorithmic reduction, this allows us to efficiently implement conversation protocols that lead to fast agreement and are guaranteed to be accuracy improving, starting with any model M (about which we make no assumptions), and any interlocutor that also satisfies conversation calibration.

Our result recovers the parameters proven by Aaronson [2005] for the special case of agreement by fully rational Bayesian forecasters in a setting with a known prior. Moreover, our result generalizes beyond the setting of Aaronson [2005] in a number of ways. For example, it straightforwardly generalizes to the setting in which the outcome $y^t \in [0,1]^d$ is d-dimensional, by requiring that the forecasts satisfy conversation calibration marginally in each coordinate. This comes at a cost of d in our convergence bounds:

Theorem 1.4 (Informal, see Theorem 5.2)**.** *When $\mathcal{Y} = [0,1]^d$ and the human and model satisfy conversation-calibration marginally in each coordinate (Definition 3.23), then for any $\epsilon, \delta \in (0,1]$ and large enough T, on a $1-\delta$ fraction of days, they reach ϵ-agreement after at most $K = \frac{d}{\epsilon^2 \delta}$ rounds of conversation. Furthermore, for this $1-\delta$ fraction of days, if they reach agreement in round i, their final predictions have a lower squared error than the base predictions of either the human or the model, by a term that scales as $\frac{i}{\delta \epsilon^2}$.*

We similarly show an efficient reduction that can convert an arbitrary model M into an algorithm capable of engaging in a conversation protocol, that is guaranteed to satisfy conversation calibration marginally in each coordinate when interacting with any interlocutor (see Theorem 5.6).

Agreeing on Actions Although our analysis of the "canonical setting" extends to d dimensional agreement, it requires that both parties provide d-dimensional numeric predictions at each round. The model in our reduction is able to do this, but it would be better not to require the human interlocutor to provide numeric feedback, especially in high dimensions when $y^t \in [0,1]^d$. To avoid this, we adopt a downstream decision-making perspective. We imagine that the human has an action set $\mathcal{A}$ (e.g. treatments and diagnostic tests that a doctor could order), as well as a utility function $U : \mathcal{A} \times [0,1]^d \to [0,1]$ that maps an action $a \in \mathcal{A}$ and a label $y \in [0,1]^d$ to a utility $U(a, y)$ that the human would like to maximize. We assume that the utility function is linear in its second argument. This captures (among other things) the scenario in which there are d discrete outcomes for which the human has arbitrary utilities, the predictions are *probability distributions* over these d outcomes, and the human is an expectation maximizer.

We define a calibration condition that we call "conversation decision calibration", which additionally conditions on the action most recently suggested by one's interlocutor. Like our definition of conversation calibration, conversation decision calibration only involves a small number $(|\mathcal{A}|^2)$ of conditioning events, and so is computationally and statistically tractable to enforce — in this case by using the online algorithm for making d dimensional forecasts unbiased subject to an arbitrary polynomial collection of conditioning events given by Noarov et al. [2023]. We provide analogous results for the action-agreement setting, in which we avoid a dependence on the dimension d. We discuss this setting in Section 6.

One-Shot Guarantees for Bayesians. Bayesian posterior beliefs, when computed from a known (and correct) prior are known to be well calibrated [Dawid, 1982]. We show that this extends to our notions of conversation calibration: when instances are drawn i.i.d. from a fixed and known prior, then a Bayesian, reporting at every round their posterior expectation for y, will satisfy our notions of decision calibration, no matter how their interlocutor is making predictions. The most immediate implication of this is that the calibration assumptions that our convergence results rely on are all strict relaxations of Bayesian rationality. However it also allows us to lift all of our convergence guarantees to the "one-shot" setting when two Bayesians with a shared prior are conversing with one another. Rather than speaking of a sequence of conversations over days and making guarantees on the maximum length of conversations on all but a $1-\delta$ fraction of days, we can make exactly the same guarantees on the length of a single conversation between two Bayesians, with probability $1-\delta$ over the draw of the instance from their commonly shared prior distribution. For example, when applied to our 1-dimensional convergence result in the canonical setting, we recover the Theorem of Aaronson [2005]: Two Bayesians will reach ϵ-agreement after $k = O(1/\epsilon^2\delta)$ many rounds with probability $1-\delta$ over the draw of the instance from the prior. Our other results generalize Aaronson [2005] and lead to new theorems about the rate of convergence in the Bayesian agreement setting – agreement on d-dimensional expectations and agreement using action feedback.

Theorem 1.5 (Informal, see Corollary 7.8)**.** *Fix any prior $\mathcal{D}$ over triples (x_h, x_m, y). For an instance drawn from $\mathcal{D}$, with probability $1-\delta$ over the draw of the instance:*

1. *In the d-dimensional "full-feedback" setting two Bayesian parties agree after exchanging at most $K \leq \frac{3d}{\epsilon^2\delta}$ messages.*

2. *In the action-feedback setting, two Bayesian parties agree after exchanging at most $K \leq \frac{3}{2\varepsilon\delta} + 1$ messages — i.e. they obtain a dimension independent convergence rate.*

Extension to n Parties. We can extend all of our results from 2 parties to n parties, with only a polynomial overhead in n in terms of computational and statistical complexity of the n parties, and no dependence on n in terms of how many times each party must speak before agreement (in total the number of "rounds" of conversation increases by a factor of n simply because it now takes n rounds in between each agent speaking two times consecutively). We discuss this extension in Section E.

References

Scott Aaronson. The complexity of agreement. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 634–643, 2005.

Rohan Alur, Manish Raghavan, and Devavrat Shah. Human expertise in algorithmic prediction. *arXiv preprint arXiv:2402.00793*, 2024.

Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth, and Mirah Shi. An elementary predictor obtaining $2\sqrt{T}+1$ distance to calibration. *IEEE Symposium on Discrete Algorithms (SODA)*, 2025. URL `https://arxiv.org/abs/2402.11410`.

Robert J. Aumann. Agreeing to Disagree. *The Annals of Statistics*, 4(6):1236 – 1239, 1976. doi: 10.1214/aos/1176343654. URL `https://doi.org/10.1214/aos/1176343654`.

Abhijit V Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107 (3):797–817, 1992.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023.

Modibo K Camara, Jason D Hartline, and Aleck Johnsen. Mechanisms for a no-regret agent: Beyond the common prior. In *2020 ieee 61st annual symposium on foundations of computer science (focs)*, pages 259–270. IEEE, 2020.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024.

Natalie Collina, Aaron Roth, and Han Shao. Efficient prior-free mechanisms for no-regret agents. *The ACM Conference on Economics and Computation (EC)*, 2024.

Yuval Dagan, Constantinos Daskalakis, Maxwell Fishelson, Noah Golowich, Robert Kleinberg, and Princewill Okoroafor. Improved bounds for calibration via stronger sign preservation games. *arXiv preprint arXiv:2406.13668*, 2024.

A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379): 605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856`.

A. P. Dawid. Calibration-Based Empirical Probability. *The Annals of Statistics*, 13(4):1251 – 1274, 1985. doi: 10.1214/aos/1176349736. URL `https://doi.org/10.1214/aos/1176349736`.

Yash Deshpande, Elchanan Mossel, and Youngtak Sohn. Agreement and statistical efficiency in bayesian perception models. *arXiv preprint arXiv:2205.11561*, 2022.

Ally Yalei Du, Dung Daniel Ngo, and Zhiwei Steven Wu. Reconciling model multiplicity for downstream decision making. *arXiv preprint arXiv:2405.19667*, 2024a.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024b.

Dean P Foster and Sham M Kakade. Calibration via regression. In *2006 IEEE Information Theory Workshop-ITW'06 Punta del Este*, pages 82–86. IEEE, 2006.

Dean P. Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1):7–35, 1999. ISSN 0899-8256. doi: https://doi.org/10.1006/game.1999.0740. URL `https://www.sciencedirect.com/science/article/pii/S0899825699907406`.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Rafael Frongillo, Eric Neyman, and Bo Waggoner. Agreement implies accuracy for substitutable signals. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 702–733, 2023.

Douglas Gale and Shachar Kariv. Bayesian learning in social networks. *Games and economic behavior*, 45(2):329–346, 2003.

Sumegha Garg, Michael P Kim, and Omer Reingold. Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 809–824, 2019.

John D Geanakoplos and Heraklis M Polemarchakis. We can't disagree forever. *Journal of Economic theory*, 28(1):192–200, 1982.

Ira Globus-Harris, Varun Gupta, Michael Kearns, and Aaron Roth. Model ensembling for constrained optimization, 2024. URL `https://arxiv.org/abs/2405.16752`.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *Innovations in Theoretical Computer Science (ITCS)*, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Lunjia Hu and Yifan Wu. Predict to minimize swap regret for all payoff-bounded tasks. *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2024.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5143–5145. PMLR, 2023.

Yuqing Kong and Grant Schoenebeck. False consensus, information theory, and prediction markets. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251, page 81. Schloss Dagstuhl–Leibniz-Zentrum f {\" u} r Informatik, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

George J Mailath and Larry Samuelson. *Repeated Games and Reputations: Long-Run Relationships*. Oxford University Press, 2006.

Elchanan Mossel, Allan Sly, and Omer Tamuz. Asymptotic learning on bayesian social networks. *Probability Theory and Related Fields*, 158(1):127–157, 2014.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making, 2023. URL `https://arxiv.org/abs/2310.17651`.

Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 456–466, 2021.

Mingda Qiao and Letian Zheng. On the distance from calibration in sequential prediction. *arXiv preprint arXiv:2402.07458*, 2024.

Aaron Roth and Mirah Shi. Forecasting for swap regret for all downstream agents. *The ACM Conference on Economics and Computation (EC)*, 2024.

Aaron Roth, Alexander Tolbert, and Scott Weinstein. Reconciling individual probability forecasts, 2023. URL `https://arxiv.org/abs/2209.01687`.

Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. Calibration with many checking rules. *Mathematics of operations Research*, 28(1):141–153, 2003.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323, 1992.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.

2 Appendix

2.1 Related Work

Agreement. Aumann's classic "agreement theorem" Aumann [1976] states that two Bayesians with a common and correct prior, who have *common knowledge* of each other's posterior expectation of any predicate must have the same posterior expectation of that predicate. This sparked a very large literature on agreement amongst Bayesians — we touch upon only the most related work here. "Common Knowledge" is the limit of an infinite exchange of information, but Geanakoplos and Polemarchakis Geanakoplos and Polemarchakis [1982] showed that whenever the underlying state space is finite, then agreement occurs after a finite number rounds (depending on the cardinality of the state space) in which the information exchanged in each round is the posterior expectation of each party. Aaronson Aaronson [2005] showed that for 1-dimensional expectations, ϵ-approximate agreement can be obtained (with probability $1-\delta$ over the draw from the prior distribution) after the parties exchange only $O(1/\epsilon^2\delta)$ messages. Notably this bound is independent of the representation size or complexity of the underlying prior distribution. Two papers Kong and Schoenebeck [2023], Frongillo et al. [2023] study conditions under which Aumannian agreement implies information aggregation — i.e. when "agreement" is reached at the same posterior belief that would have resulted had the two parties shared all of their information, rather than interacting within an agreement protocol. There is also a large literature that studies multi-party agreement amongst Bayesians connected via a communication network — see e.g. Geanakoplos and Polemarchakis [1982], Gale and Kariv [2003], Aaronson [2005], Mossel et al. [2014], Deshpande et al. [2022]. Part of this literature (e.g. Gale and Kariv [2003], Mossel et al. [2014]) studies settings in which the beliefs of the parties are not directly observed, but rather what action they take is, under the presumption that they take a utility maximizing action. In general this literature is interested in exact asymptotic agreement. These papers also all assume that there is a commonly known prior and that all parties are able to compute correct posterior expectations for the predicate of interest, despite the fact that this might be computationally intractable. Aaronson Aaronson [2005] gives a computational reduction from the problem of participating in an agreement protocol to the problem of computing (and sampling from) correct posterior distributions, conditional on any vector of features that might be observed by either party. This might itself be a computationally hard task, and the reduction requires a number of calls to this posterior-computation oracle that is super-exponential in $1/\epsilon$, but independent of the cardinality of the state space. Our primary point of departure from this literature is that we ask for algorithms that are truly computationally tractable (i.e. worst-case polynomial time in all parameters) and make no distributional assumptions, although when there is a commonly known prior and agents are Bayesians, we recover theorems in the classical Aumannian setting. This leads to new quantitative agreement theorems in the setting and style of Aaronson [2005] (i.e. bounds that depend only on approximation parameters and are independent of the complexity of the instance) — in d dimensional settings. In particular, that agents providing only action feedback in d-dimensional belief spaces will arrive at ϵ-agreement after only $O(1/\epsilon\delta)$ many rounds of conversation with probability $1-\delta$ (independent of the dimension d).

Calibration. Our techniques are rooted in the ability to maintain *calibrated* forecasts in online adversarial settings, which was first shown by Foster and Vohra Foster and Vohra [1998]. Calibration itself dates back to Dawid Dawid [1982, 1985], who also showed that Bayesians with correctly specified priors are calibrated Dawid [1982]. Conditional calibration guarantees also have a long history Dawid [1985], Sandroni et al. [2003], Foster and Kakade [2006] with a recent seminal formalization as *multicalibration* Hébert-Johnson et al. [2018] which can be obtained with good rates in both the batch and online adversarial settings Gupta et al. [2022]. The traditional calibration measure of "expected calibration error" has a number of shortcomings; the most relevant for us is that it cannot be obtained with $O(\sqrt{T})$ rates in online adversarial settings Qiao and Valiant [2021], Dagan et al. [2024]. This has led to a recent exploration of alternative calibration measures, notably "distance to calibration" Błasiok et al. [2023]. Distance to calibration *can* be obtained at $O(\sqrt{T})$ rates in online adversarial settings with extremely simple, deterministic algorithms Qiao and Zheng [2024], Arunachaleswaran et al. [2025], and turns out to be sufficient for our application. In particular our reduction in the canonical case uses the algorithm of Arunachaleswaran et al. [2025]. In our "action feedback" setting we use a variant of "decision calibration" Zhao et al. [2021], which can similarly be guaranteed in online adversarial settings with good rates, using the algorithm of Noarov et al. [2023]. This is related to a line of recent work exploring notions of calibration tailored to downstream decision-making Kleinberg et al. [2023], Noarov et al. [2023], Roth and Shi [2024], Hu and Wu [2024].

Several papers Camara et al. [2020], Collina et al. [2024] have replaced traditional assumptions of Bayesian rationality (and common prior assumptions) with calibration assumptions in *principal agent* problems arising e.g. in contract theory and Bayesian Persuasion. In particular, Collina et al. [2024] shows how to do this with tractable decision calibration conditions. Beyond this, the most thematically related use of calibration is its use as an ensembling method. Garg et al. [2019] shows how to produce a model that is "cross calibrated" to two models, and is more accurate than each while improving various fairness measures of decisions downstream of the model. Roth et al. [2023] shows how to use cross-calibration to resolve "predictive multiplicity", and derive a single more accurate model from any pair of models that are equally accurate and yet frequently disagree. This kind of ensembling was recently extended to agreement for downstream actions Du et al. [2024a] and for ensembling models for high dimensional downstream optimization problems Globus-Harris et al. [2024]. Alur, Raghavan, and Shah employ a similar model ensembling approach motivated by human-AI collaboration Alur et al. [2024]. Informally, they learn a model in a batch setting that is cross-calibrated to the fixed judgments of a human. What distinguishes our work from the line of work using calibration for ensembling (aside from the fact that we work in the online adversarial setting) is that prior work in this area treats the models to be ensembled as static. That is, the models to be ensembled are defined by fixed mappings from features to predictions, and do not update their beliefs as a function of interaction with other models. As a result, these methods cannot be applied to Bayesian-like entities which requires the kind of interactive conversation protocol we adopt in this work.

Multi-agent Debates with LLMs. To improve the accuracy of the responses of large language model (LLM) generations, recent work Du et al. [2024b], Liang et al. [2023], Chan et al. [2024] has proposed the multi-agent debate approach in which two (or more) LLMs "debate" their individual responses and reasoning processes in multiple rounds until they converge to a final answer, and then a "judge" (often another LLM or human) validates the final answer. Here, debate loosely refers to the two LLMs getting to see each other's responses after each round and update their subsequent responses. This shares notable similarities to our agreement protocol where the LLMs map to the agents, the messages map to the generations of each LLM in each round, and the judge maps to the outcome label at each day. Moreover, they share the general motivation to improve the accuracy of the agents using interaction and the assumption that agents behave in good faith. In contrast to our work which deals with numerical predictions and makes formal calibration assumptions on the agents, multi-agent debates operate in natural language under less formal assumptions. We believe our framework of agreement protocols could potentially be adapted to analyze multi-agent LLM debate dynamics and explain why LLMs reach consensus and improve overall accuracy. Additionally, our techniques to enforce the calibration conditions could be useful to improve the efficiency and performance of LLM debates. We note that prior work Deshpande et al. [2022] has also discussed viewing communicating LLMs through the lens of agreement. However, their work assumes that LLMs are purely Bayesian agents with a shared prior, and they focus only on statistical efficiency of reaching agreement on a network, not the rate of convergence to agreement.

3 Preliminaries

In most of the paper we study a setting with two agents, whom we call the *human* and the *model* (in Appendix E we generalize our results the setting to $n \geq 2$ agents). Both the human and model are able to make predictions about a label not just in isolation (given features), but as a function of an interaction that they have had with another agent. The agents interact to make predictions over a sequence of days $t = 1, \ldots, T$. We let $\mathcal{X}_h$ and $\mathcal{X}_m$ denote feature spaces for the human and the model respectively. We let $\mathcal{Y}$ represent the outcome (label) space, which we always take to be real or vector valued, so that we can sensibly speak of expectations over it. For notational simplicity, we will assume that $\mathcal{Y}$ is convex, so that expectations over $\mathcal{Y}$ are themselves elements of $\mathcal{Y}$, although this is not necessary.

On each day t, the human and model aim to reach agreement, with respect to some agreement condition, on their predictions of that day's outcome y^t based on the features they each see: x_h^t and x_m^t, respectively. They do so by conversing over a series of *rounds* $k = 1 \ldots, L$. The human and model will alternate speaking, and we suppose that the model acts in odd numbered rounds; the human acts in even numbered rounds. In an odd round k, the model sends a message $p_m^{t,k}$, and then in the next round $k+1$, the human responds with a message $p_h^{t,k+1}$. We write Ω_h for the message space of the human and Ω_m for the message space of the model. At each round k when they are speaking, an agent has an underlying prediction of the (expectation of the) label, denoted $\hat{y}_m^{t,k}$ and $\hat{y}_h^{t,k}$ respectively. This underlying prediction can be a function of everything the agent has observed so far — the features relevant to the instance, the messages sent by the other party, and past outcomes on previous days. The message each agent sends at each round will be a function of this underlying prediction. For example, the messages sent might be the underlying predictions themselves (as in the full feedback setting we study in Sections 4 and 5) — but the messages might also be some "coarsening" of the prediction, as in the action feedback setting we study in Section 6. The day terminates once an agreement condition is met, at which point that day's label y^t is revealed to both parties.

3.1 Agreement Protocols

We study a variety of settings, each of which is instantiated by the label space, the message space of each of the parties, and an agreement condition. We begin by defining a generic agreement condition, which we can instantiate for each particular setting. Informally, the agreement condition takes in the messages and underlying predictions of each agent, and decides if they are sufficiently close to terminate the conversation for that day.

Definition 3.1 (Agreement Condition). *An agreement condition is a function that determines when the human and model's predictions are "ε-close", for any $\varepsilon > 0$ as a function of their most recently sent messages and predictions:* $\textsc{Agree}_\varepsilon : (\Omega_h, \mathcal{Y}) \times (\Omega_m, \mathcal{Y}) \to \{0, 1\}$. *An agreement condition should be the conjunction of two conditions (one for the model and one for the human), each of which can be evaluated with only knowledge of* their own *predictions and their counter-party's message. In other words we should be able to write* $\textsc{Agree}_\varepsilon(p_h, \hat{y}_h, p_m, \hat{y}_m) = \textsc{Agree}_\varepsilon^h(p_m, p_h, \hat{y}_h) \cdot \textsc{Agree}_\varepsilon^m(p_m, p_h, \hat{y}_m)$ *for some pair of functions* $\textsc{Agree}_\varepsilon^h, \textsc{Agree}_\varepsilon^m$.

Remark 3.2. *In practice, whether each agent is in agreement with the other is determined by the agent—this is why we want agreement conditions to be the conjunction of a pair of conditions each of which can be evaluated by each agent in isolation. The formalism of an "agreement condition" is only to let us easily describe and instantiate our various settings.*

As an example, we can consider the simplest agreement condition we use. This is the agreement condition for the full feedback, one-dimensional prediction ("canonical") setting: when $\Omega_h = \Omega_m = \mathcal{Y} = [0, 1]$.

Definition 3.3 (Agreement Condition in the Canonical Setting). *The agreement condition in the canonical setting is the function* $\textsc{Agree-Canonical}_\varepsilon : \Omega_h \times \mathcal{Y} \times \Omega_m \times \mathcal{Y} \to \{0, 1\}$ *defined as:*

$$\textsc{Agree}_\varepsilon(p, y_h, q, y_m) = \begin{cases} 1, & \textit{if } |p - q| < \varepsilon \\ 0, & \textit{otherwise.} \end{cases}$$

We formalize the interaction between the two agents in Protocol 3.1 — a generic "agreement protocol"—which can be instantiated with the particulars of each setting we study.

[ht]

```
Input $(\Omega_h, \Omega_m, \mathcal{Y}, \textsc{Agree}_\epsilon)$
for each day $t = 1, \ldots$ do
  Receive $x^t = (x^t_h, x^t_m)$. The model sees $x^t_m$ and the human sees $x^t_h$.
  for each round $k = 1, 2, \ldots, L$ do
    if $k$ is odd then
      The Model predicts $\hat{y}^{t,k}_m \in \mathcal{Y}$, and sends the Human $p^{t,k}_m \in \Omega_m$
      if $\textsc{Agree}_\varepsilon(p^{t,k-1}_h, \hat{y}^{t,k-1}_h, p^{t,k}_m, \hat{y}^{t,k}_m)$ then
        Return $p^{t,k}_m$ and break out of loop
    if $k$ is even then
      The Human predicts $\hat{y}^{t,k}_h$, and sends the model $p^{t,k}_h \in \Omega_h$
      if $\textsc{Agree}_\varepsilon(p^{t,k}_h, \hat{y}^{t,k}_h, p^{t,k-1}_m, \hat{y}^{t,k-1}_m)$ then
        Return $p^{t,k-1}_m$ and break out of loop
  The Human and Model observe $y^t \in \mathcal{Y}$
```

When interacting within an agreement protocol, we say that on day t the two agents agree after k rounds of conversation if the agreement condition is met at round k of day t.

3.1.1 Instantiating Different Feedback Models

We can now formally specify the various settings we study. These will vary in the label space, the message space for each participant, the mapping between predictions and messages, and the agreement condition.

Full Feedback The first setting we study is the full feedback, one-dimensional prediction, or "canonical", setting. Here, the human and model are both communicating their precise point predictions for the (expectation of the) unknown label to each other. Agreement will refer to when the human and model's predictions are sufficiently close numerically.

Definition 3.4 (Canonical Setting)**.** *The canonical setting refers to Protocol 3.1 instantiated with* $\Omega_m = \Omega_h = \mathcal{Y} = [0, 1]$*, messages* $p^{t,k}_m = \hat{y}^{t,k}_m$ *and* $p^{t,k}_h = \hat{y}^{t,k}_h$*, and the agreement condition* $\textsc{Agree}_\varepsilon = \textsc{Agree-Canonical}_\varepsilon$ *(Definition 3.3).*

This naturally extends to the d-dimensional setting, in which we measure agreement using the ℓ_∞ norm:

Definition 3.5 (Agreement Condition in the d-dimensional Setting)**.** *The agreement condition in the* $d-$*dimensional setting is the function* $\textsc{Agree-dDim}_\varepsilon : \Omega_h \times \mathcal{Y} \times \Omega_m \times \mathcal{Y} \to \{0, 1\}$ *defined as:*

$$\textsc{Agree-dDim}_\varepsilon(p, y_h, q, y_m) = \begin{cases} 1, & \textit{if } \|p - q\|_\infty < \varepsilon \\ 0, & \textit{otherwise.} \end{cases}$$

Definition 3.6 (d-dimensional Full Feedback Setting)**.** *The* d*-dimensional full feedback setting refers to Protocol 3.1 instantiated with* $\Omega_h = \Omega_m = \mathcal{Y} = [0, 1]^d$*, messages* $p^{t,k}_m = \hat{y}^{t,k}_m$ *and* $p^{t,k}_h = \hat{y}^{t,k}_h$*, and the agreement condition* $\textsc{Agree}_\varepsilon = \textsc{Agree-dDim}_\varepsilon$ *(Definition 3.5).*

Action Feedback In this setting, we study a human and a model who aim to agree on an action to take when their predictions are used to inform downstream decision-making. We model the human as having a known action set $\mathcal{A}$ and utility function $U : \mathcal{A} \times \mathcal{Y} \to [0, 1]$. The human and model are both maintaining predictions of the underlying state – which is here a d-dimensional vector — $\mathcal{Y} = [0, 1]^d$ — and are using their predictions to choose an action that is utility maximizing given the forecast. In this setting, the human and model do not exchange their estimates of the state directly, but instead simply suggest actions to one another (utility maximizing actions under their forecasts): $\Omega_h = \Omega_m = \mathcal{A}$. Here, our notion of ϵ-agreement will be that both parties agree that the action suggested by the other party obtains utility that is within ϵ of the best-response action, as measured under their own forecasts.

Definition 3.7 (Agreement Condition in the Action Feedback Setting)**.** *The agreement condition in the action feedback setting is the function* $\textsc{Agree-Action}_\varepsilon : \Omega_h \times \mathcal{Y} \times \Omega_m \times \mathcal{Y} \to \{0, 1\}$ *defined as:*

$$\textsc{Agree-Action}_\varepsilon(p, y_h, q, y_m) = \begin{cases} 1, & \textit{if } U(p, y_m) \geq U(q, y_m) - \epsilon \textit{ and } U(q, y_h)) \geq U(p, y_h) - \epsilon \\ 0, & \textit{otherwise.} \end{cases}$$

Definition 3.8 (Action Feedback Setting). *The action feedback setting refers to Protocol 3.1 instantiated with $\Omega_h = \Omega_m = \mathcal{A}$, $\mathcal{Y} = [0,1]^d$, messages $p_m^{t,k} = \arg\max_{a \in \mathcal{A}} U(a, \hat{y}_m^{t,k})$ and $p_h^{t,k} = \arg\max_{a \in \mathcal{A}} U(a, \hat{y}_h^{t,k})$, and the agreement condition* $\textsc{Agree}_\varepsilon = \textsc{Agree-Action}_\varepsilon$ *(Definition 3.7).*

Here we state the necessary assumptions for the utility functions our theorems will apply to, following the formalism of Noarov et al. [2023]:

Assumption 1 (Utility $U(\cdot,\cdot)$). *The utility function $U : \mathcal{A} \times \mathcal{Y} \to [0,1]$ maps an action a and a vector valued outcome y to a real number $U(a,y)$. We assume that for every action $a \in \mathcal{A}$:*

- *$U(a,\cdot)$ is linear in its second argument: for all $\alpha_1, \alpha_2 \in \mathbb{R}$, $y_1, y_2 \in \mathbb{R}^d$,*

$$U(a, \alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 U(a, y_1) + \alpha_2 U(a, y_2)$$

- *$U(a,\cdot)$ is L-lipschitz in its second argument in the L1-norm: for all y_1, $y_2 \in \mathbb{R}^d$,*

$$|U(a, y_1) - U(a, y_2)| \leq L \|y_1 - y_2\|_1.$$

Remark 3.9. *One natural special case is when y represents a probability distribution over d discrete outcomes $c_1, \ldots, c_d$, such that there is an arbitrary mapping $M(a,c)$ from action/outcome pairs to utilities $[0,1]$. In this case, $U(a,y)$ represents the expected utility of the action a over the outcome distribution, which is linear in y by the linearity of expectation. The utility function is L-Lipschitz in the L^1-norm, where $L = \max_{a,c_1,c_2}(M(a,c_1) - M(a,c_2))$. So this class of utility functions naturally captures any risk neutral decision maker with d payoff relevant states, but is more general.*

3.2 Algorithms for Interaction

An agreement protocol as we have defined it is used by two agents who are able to update their predictions not only as a function of the features they have observed, but as a function of an interaction with another agent. We will want to convert static models (which map features to predictions) into such interactive algorithms. In order to define such algorithms, it will be useful to establish a notation that refers to different pieces of information that both parties will have available to them at different times in Protocol 3.1, which they can use in their predictions.

We refer to the history of interaction *within* any given day t as a "conversation." This is, informally, the sequence of messages exchanged by the human and the model specifically about the currently unknown label y^t. Recall that the model and human speak in alternating (odd and even numbered, respectively) rounds.

Definition 3.10 (Conversation C). *A conversation between the human and model on day t over rounds 1 to ℓ is denoted by $C^{t,1:\ell} \in \{\Omega_m \cup \Omega_h\}^\ell$, is a sequence of ℓ messages:*

$$C^{t,1:\ell} := \begin{cases} (p_m^{t,1}, p_h^{t,2}, p_m^{t,3}, p_h^{t,4}, \ldots, p_m^{t,\ell}) & \text{if } \ell \text{ is odd,} \\ (p_m^{t,1}, p_h^{t,2}, p_m^{t,3}, p_h^{t,4}, \ldots, p_h^{t,\ell}) & \text{otherwise.} \end{cases}$$

We refer to the full conversation at day t as C^t. We define $\mathcal{C}^\ell$ to be the space of all possible conversations of length ℓ and $\mathcal{C} = \bigcup_{\ell > 0} \mathcal{C}^\ell$ represent all possible conversations.

Definition 3.11 (Conversation Length). *We define C^t to be the conversation at day t and ℓ^t to be the length of $C^t : \ell^t = |C^t|$.*

It will often be useful to consider subsequences of our objects—messages, predictions, and labels—within a certain round. We provide notation for this below.

Definition 3.12 (Round Subsequence). *For a fixed round k, we define $T^{\geq k}$ to be the subsequence of days on which conversation reaches round k, that is, $T^{\geq k} := \{t \in \{1, \ldots, T\} \mid \ell^t \geq k\}$.*

Definition 3.13 (Message Subsequence $p_m^{S,k}$). *For some set $S \subseteq \{1, \ldots, T\}$, we define $p_m^{S,k}$ as $\{p_m^{t,k} : t \in S \cap T^{\geq k}\}$, the subsequence of model predictions at round k corresponding to the subsequence of days t which reach round k and which are in the set S. We will similarly use the notation $p_h^{S,k}$, $\hat{y}_m^{S,k}$, and $\hat{y}_h^{S,k}$, to refer, respectively, to the human messages, model predictions, and human predictions over subsequences constrained in this way.*

We refer to the history of interaction *across* multiple days as a "message transcript." It is an object that records the interactions between the agents and is visible to both, and which they can use to make their predictions (unlike the "prediction transcript" which we will define immediately following).

Definition 3.14 (Message Transcript $\mu^{1:T}$). *A message transcript $\mu^{1:T} \in \{\mathcal{C} \times \mathcal{Y}\}^T$ is a sequence of conversation, outcome pairs over T days:*

$$\mu^{1:T} = \left[(C^1, y^1), \ldots, (C^T, y^T)\right].$$

We define $\mathcal{M}^T$ to be the space of all possible message transcripts over T days and $\mathcal{M} = \bigcup_{T>0} \mathcal{M}^T$ to be the space of all possible message transcripts.

We define $\mu^{t,:}$ to be the restriction of the message transcript to the elements relevant to day t — this is simply the record of the conversation at day t paired with the outcome at day t:

$$\mu^{t,:} = \begin{cases} \left((p_m^{t,1}, p_h^{t,2}, \ldots, p_m^{t,\ell^t}), y^t\right) & \text{if } \ell^t \text{ is odd,} \\ \left((p_m^{t,1}, p_h^{t,2}, \ldots, p_h^{t,\ell^t}), y^t\right) & \text{otherwise.} \end{cases}$$

Similarly, we define $\mu^{:,k}$ to be the restriction of the message transcript to the elements relevant to only round k of conversation across the subsequence of days that reach round k ($T^{\geq k}$):

$$\mu^{:,k} = \begin{cases} \left[(p_m^{t,k}, y^t) \;\middle|\; t \in T^{\geq k}\right] & \text{if } k \text{ is odd,} \\ \left[(p_h^{t,k}, y^t) \;\middle|\; t \in T^{\geq k}\right] & \text{otherwise.} \end{cases}$$

It will also be useful to be able to refer to the sequence of predictions made by the human and model across particular days or rounds. Note that depending on the setting we are working in, this "prediction transcript" will not generally be visible to both players (each player always observes their own predictions, but only the messages sent by the other):

Definition 3.15 (Prediction Transcript $\pi^{1:T}$). *A prediction transcript $\pi^{1:T} \in \left\{\bigcup_{\ell>0} (\mathcal{Y})^\ell \times \mathcal{Y}\right\}^T$ is a sequence of tuples of predictions over rounds made by the model and human (alternating across rounds), and the outcome, over T days:*

$$\pi^{1:T} = \left[\left(\hat{y}_m^{1,1}, \hat{y}_h^{1,2}, \hat{y}_m^{1,3}, \ldots \hat{y}_m^{1,\ell_1}, y^1\right), , \ldots, \left(\hat{y}_m^{T,1}, \hat{y}_h^{T,2}, \hat{y}_m^{T,3}, \ldots \hat{y}_m^{T,\ell^T}, y^T\right)\right]$$

Similar to Definition 3.14, we define $\pi^{t,:}$ to be the restriction to elements relevant to day t and $\pi^{:,k}$ to be the restriction to only round k of conversation across days as follows:

$$\pi^{t,:} = \begin{cases} \left((\hat{y}_m^{t,1}, \hat{y}_h^{t,2}, \hat{y}_m^{t,3}, \ldots, \hat{y}_m^{t,\ell_t}), y^t)\right) & \text{if } \ell^t \text{ is odd,} \\ \left((\hat{y}_m^{t,1}, \hat{y}_h^{t,2}, \hat{y}_m^{t,3}, \ldots, \hat{y}_h^{t,\ell_t}), y^t)\right) & \text{otherwise.} \end{cases} \qquad \pi^{:,k} = \begin{cases} \left[(\hat{y}_m^{t,k}, y^t) \;\middle|\; t \in T^{\geq k}\right] & \text{if } k \text{ is odd,} \\ \left[(\hat{y}_h^{t,k}, y^t) \;\middle|\; t \in T^{\geq k}\right] & \text{otherwise.} \end{cases}$$

Finally, we will also use the restriction $\pi_h^{1:T}$ and $\pi_m^{1:T}$ the prediction transcript restricted to the human and model predictions, respectively, through day T.

With these definitions in hand, we can now give a formal specification of the types of algorithms we will be using in our results.

Definition 3.16 (Model Algorithm M). *The Model's algorithm $M : \mathcal{M} \times \Pi \times \mathcal{C} \times \mathcal{X}_m \to \Delta\Omega_m$ is a mapping from a t-length message transcript, a prediction transcript of the model's predictions through round t $\pi_m^{1:t}$, an ℓ-length conversation, and a feature vector x_m^{t+1} to a distribution over messages $p_m^{t+1,\ell+1}$ for day $t+1$ in round $\ell+1$.*

Definition 3.17 (Human Algorithm H). *The Human's algorithm $H : \mathcal{M} \times \Pi \times \mathcal{C} \times \mathcal{X}_h \to \Delta\Omega_h$ is a mapping from a t-length message transcript, a prediction transcript of the humans's predictions through round t $\pi_h^{1:t}$, an ℓ-length conversation, and a feature vector x_h^{t+1} to a distribution over messages $p_h^{t+1,\ell+1}$ for day $t+1$ in round $\ell+1$.*

3.3 Calibration

The main focus of our work is studying computationally tractable conditions under which the two parties achieve fast agreement in the models described in Section 3.1.1. The conditions that we study (and enforce) will be *calibration* conditions of various sorts. In this section we give the basic calibration definitions that we will be working with.

The standard measure of calibration of some sequence of predictions $p^{1:T}$ to outcomes $y^{1:T}$ in a sequential prediction setting is *expected calibration error*, defined as follows.

Definition 3.18 (Expected Calibration Error)**.** *Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, their expected calibration error is,*

$$\mathrm{ECE}(p^{1:T}, y^{1:T}) = \sum_{p\in[0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|$$

Here the outer sum is over the values p that appear in the sequence $p^{1:T}$.

We will sometimes measure calibration error of a sequence instead using *distance to calibration*, first defined by Błasiok et al. [2023] (we here use the definition given by Qiao and Zheng [2024] in the sequential setting). Distance to calibration measures the ℓ_1 distance between a sequence of predictions and the closest sequence of *perfectly calibrated* predictions.

Definition 3.19 (Distance to Calibration)**.** *Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, the distance to calibration is,*

$$\mathrm{CalDist}(p^{1:T}, y^{1:T}) = \min_{q^{1:T}\in\mathcal{C}(y^{1:T})} \left\| p^{1:T} - q^{1:T} \right\|_1$$

where $\mathcal{C}(y^{1:T}) = \{q^{1:T} : \mathrm{ECE}(q^{1:T}, y^{1:T}) = 0\}$ is the set of predictions that are perfectly calibrated against outcomes $y^{1:T}$.

Calibration has a close relationship to squared error, which we will use as a potential function in some of our analyses. Below we define the squared error of a sequence of predictions relative to a sequence of outcomes:

Definition 3.20 (Squared Error)**.** *Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, the squared error between them is,*

$$\mathrm{SQE}(p^{1:T}, y^{1:T}) := \sum_{t\in[T]} (p^t - y^t)^2.$$

We will overload this notation for the special case of constant sequences $p^1 = \ldots = p^T = p$:

$$\mathrm{SQE}(p, y^{1:T}) := \sum_{t\in[T]} (p - y^t)^2.$$

3.3.1 Conversation Calibration

We now define a new notion of calibration that we will make use of in the "canonical" setting, that we call *conversation calibration*. Informally, an agent is *conversation calibrated* if for every round of conversation k, the sequence of predictions (over days t) that they make at round k of conversation is calibrated not just marginally, but *conditionally* on the value of the prediction that the other agent made at round $k-1$. In fact, without making assumptions on the other agent, it will not be possible to give calibration guarantees that hold conditional on their predictions, because these may come from an arbitrarily large range. So instead we will condition on *bucketings* of their predictions.

Definition 3.21 (Bucketing of the Prediction Space)**.** *For bucket coarseness parameter n, let $B_n(i) = \left[\frac{i-1}{n}, \frac{i}{n}\right)$ and $B_n(n) = \left[\frac{n-1}{n}, 1\right]$ form a set $\mathcal{B}_n$ of n buckets of width $1/n$ that partition the unit interval.*

Next we define conversation calibration, which allows for calibration error as measured using distance to calibration.

Definition 3.22 (Conversation-Calibrated Predictions)**.** *Fix an error function $f : \{1, \ldots, T\} \to \mathbb{R}$ and bucketing function $g : \{1, \ldots, T\} \to (0, 1]$. Given a prediction transcript $\pi^{1:T}$ resulting from an interaction in the canonical setting (Definition 3.4), a human is (f, g)-conversation-calibrated if for all even rounds k and buckets $i \in \{1, \ldots, 1/g(T)\}$:*

$$\mathrm{CalDist}(p_h^{T_m(k,i),k}, y^{T_m(k,i)}) \leq f(|T_m(k, i)|),$$

where $T_m(k,i) = \left\{t \in T^{\geq k} \mid p_m^{t,k-1} \in B_i(1/g(T))\right\}$ is the subsequence of days where the predictions of the model at the previous round fall in bucket i and the conversation reaches round k.

Symmetrically, a model is (f, g)-conversation-calibrated if for all odd rounds k and buckets $i \in \{1, \ldots, 1/g(T)\}$:

$$\mathrm{CalDist}(p_m^{T_h(k,i),k}, y^{T_h(k,i)}) \leq f(|T_h(k, i)|),$$

where $T_h(k, i) = \{t \in T^{\geq k} \mid p_h^{t,k-1} \in B_i(1/g(T))\}$, that is, the subsequence of days where the predictions of the human in the previous round fall in bucket i and the conversation reaches round k.

When convenient we will assume that $f(\cdot)$ is concave. This captures the case where $f(T) = T^\alpha$ for any $\alpha \in [0, 1]$, which is the form that all calibration bounds we are aware of take.

Assumption 2. *$f(\cdot)$ is a concave function.*

We also define a d-dimensional notion of conversation calibration. A naive (and intractable) generalization of conversation calibration would require that an agent's d-dimensional forecasts be (fully) calibrated conditional on the value of the d-dimensional forecasts made at the previous round by the other agent. But this would require making predictions that are unbiased subject to an exponential (in d) number of conditioning events. Instead our generalization requires that the forecasts made by each party satisfy a *marginal* conversation calibration condition in each coordinate of their prediction. That is, each coordinate i of an agent's prediction should be calibrated marginally, conditional on the value of the other agent's previous prediction *in coordinate i*. This increases the number of conditioning events compared to the 1 dimensional case only by a factor of d, and hence will be tractably obtainable.

Definition 3.23 (Conversation-Calibrated Vector Predictions)**.** *Fix an error function $f : \{1, \ldots, T\} \to \mathbb{R}$ and bucketing function $g : \{1, \ldots, T\} \to (0, 1]$. Given a prediction transcript $\pi^{1:T}$ resulting from an interaction in the full-feedback, d-dimensional setting (Definition 3.6), a human is (f, g)-conversation-calibrated if for all even rounds k, indices $j \in [d]$, and buckets $i \in \{1, \ldots, 1/g(T)\}$:*

$$\mathrm{CalDist}(p_h^{T_m(k,i,j),k}[j], y^{T_m(k,i,j)}[j]) \leq f(|T_m(k, i, j)|),$$

where $T_m(k, i, j) = \left\{t \in T^{\geq k} \mid p_m^{t,k-1}[j] \in B_i(1/g(T))\right\}$ is the subsequence of days where the j'th coordinate of the predictions of the model at the previous round fall in bucket i and the conversation reaches round k.

Symmetrically, a model is (f, g)-conversation-calibrated if for all odd rounds k, indices $j \in [d]$, and buckets $i \in \{1, \ldots, 1/g(T)\}$:

$$\mathrm{CalDist}(p_m^{T_h(k,i,j),k}[j], y^{T_h(k,i,j)}[j]) \leq f(|T_h(k, i, j)|),$$

where $T_h(k, i, j) = \{t \in T^{\geq k} \mid p_h^{t,k-1}[j] \in B_i(1/g(T))\}$.

3.3.2 Decision Conversation Calibration

Next we turn to the action feedback setting (Definition 3.8). The outcome space $\mathcal{Y}$ is now vector valued, and instead of communicating vector valued predictions as messages, the agents communicate downstream *actions*. We define decision-conversation-calibration, which asks for "decision calibration" Zhao et al. [2021], Noarov et al. [2023] conditional on the previous message sent by the other agent. In other words, the predictions that each agent makes should be unbiased conditional on both 1) the best response action implied by the predictions themselves, and 2) the best response action communicated at the previous round. Here we use an expected-calibration-error style definition, since this is what we can achieve algorithmically using the algorithm of Noarov et al. [2023].

Definition 3.24 (Decision-Conversation-Calibrated (DC-Calibrated) Predictions)**.** *Given a prediction transcript $\pi^{1:T}$ resulting from an interaction in the action feedback setting (Definition 3.8), a human is $f(\cdot)$-decision-conversation-calibrated (or $f(\cdot)$-DC-calibrated) if for all even rounds k, coordinates $i \in [d]$, and pairs of actions $a, a' \in \mathcal{A}$:*

$$\left| \sum_{t=1}^{T} \mathbb{1}[t \in T_h(k, a, a')](\hat{y}_h^{t,k}[i] - y^t[i]) \right| \leq f(|T_h(k, a, a')|),$$

where $T_h(k, a, a') = \{t \in T^{\geq k} \mid p_m^{t,k-1} = a \text{ and } p_h^{t,k} = a'\}$ is the subsequence of days in which the model's recommendation on round $k-1$ is a and the human's recommendation on round k is a'.

Symmetrically, a model is $f(\cdot)$-DC-calibrated if for all odd rounds k, coordinates $i \in [d]$, and pairs of actions $a, a' \in \mathcal{A}$:

$$\left| \sum_{t=1}^{T} \mathbb{1}[t \in T_m(k, a, a')](\hat{y}_m^{t,k}[i] - y^t[i]) \right| \leq f(|T_m(k, a, a')|),$$

where $T_m(k, a, a') = \{t \in T^{\geq k} \mid p_h^{t,k-1} = a \text{ and } p_m^{t,k} = a'\}$ is the subsequence of days in which the human's recommendation on round $k-1$ is a and the model's recommendation on round k is a'.

4 Agreement in the Canonical Setting

In this section we study the simple "canonical" setting (Definition 3.4) in which $\mathcal{Y} = \Omega_m = \Omega_h = [0, 1]$, which most closely maps onto the relevant prior work stemming from Aumann's agreement theorem Aumann [1976], Geanakoplos and Polemarchakis [1982], Aaronson [2005], Frongillo et al. [2023]. We show that when interacting in the Agreement protocol (Protocol 3.1), if both agents satisfy appropriately instantiated *conversation calibration* conditions (Definition 3.22), then once the total number of days T is sufficiently large, on a $1-\delta$ fraction of days, they ϵ-agree after at most $K \leq 2/(\epsilon^2\delta)$ rounds of conversation without reducing accuracy. We give an efficient reduction through which any static model can be converted into an algorithm satisfying these conversation calibration conditions after at most $T \leq O\left(\frac{1}{\epsilon^6\delta^3}\right)$ days. We remark that this bound is possible because we are able to carry out our analysis using *distance to calibration* bounds, which admit algorithms that obtain $O(\sqrt{T})$ rates in online adversarial settings Qiao and Zheng [2024], Arunachaleswaran et al. [2025] — we would obtain worse rates if we used the same reduction using algorithms bounding expected calibration error Qiao and Valiant [2021].

As predictions are the same as messages in the canonical setting ($p_m^{k,t} = \hat{y}_m^{k,t}$ and $p_h^{k,t} = \hat{y}_h^{k,t}$), in this section we will refer to both these terms as $p_m^{k,t}$ (and $p_h^{k,t}$) for simplicity. The following theorem formalizes the statement that conversation calibration (at sufficiently diminishing rates) guarantees fast agreement on most rounds, and that the resulting conversations improve accuracy.

Theorem 4.1. *If the Human is (f_h, g_h)-conversation-calibrated and the Model is (f_m, g_m)-conversation-calibrated, then for any $\epsilon, \delta \in [0, 1]$, on a $1-\delta$ fraction of days, they reach ϵ-agreement after at most K rounds of conversation for*

$$K \leq \frac{1}{\epsilon^2\delta - \beta(T)}$$

where $\beta(T) = 3\left(g_m(T) + g_h(T) + \frac{f_m(g_m(T)\cdot T)}{g_m(T)\cdot T} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)\cdot T}\right)$, a term that will tend to 0 for appropriately instantiated functions g and f.

Furthermore, for any round k such that $|T^{\geq k}| \geq \delta T$, we have that

$$\frac{\text{SQErr}(p_h^{T^{\geq k},k}, y_h^{T^{\geq k},k})}{T} \leq \frac{\text{SQErr}(p_h^{T^{\geq k},2}, y_h^{T^{\geq k},k})}{T} - k(\epsilon^2\delta - \beta(T)).$$

In other words, each round of conversation is error improving compared to the initial predictions of the human (or the model), with the error improving at a rate that is linear in the number of rounds of conversation.

A corollary of this theorem is that after T is taken to be sufficiently large, agreement occurs rapidly on almost every day, and each further round of conversation leads to an $\epsilon^2\delta$ decrease in squared error.

Corollary 4.2. *When $\beta(T) \leq \frac{\delta\epsilon^2}{2}$, on a $1-\delta$ fraction of days, the number of rounds until agreement is at most $K \leq \frac{2}{\delta\epsilon^2}$.*

Finally, in Theorem 4.8 we give a reduction that allows us to convert an arbitrary model into an algorithm that satisfies $(\sqrt{T}, T^{\frac{-1}{3}})$-conversation calibration, for which it suffices to take $T \geq O(\frac{1}{\epsilon^6\delta^3})$ to satisfy the conditions of Corollary 4.2.

We now turn to proving Theorem 4.1. First we give some intuition for the theorem. Our analysis will focus on the sequence of predictions made at each *round k* of conversation, over all days for which the conversation reaches that round. Intuitively, there are two cases:

1. In the first case, on most days, the prediction at round k is within ϵ of the prediction made at round $k-1$. In this case, most conversations that make it to round k end in agreement at round k.

2. In the second case, most predictions at round k differ by more than ϵ from the predictions at round $k-1$. But the sequence of predictions made at round k satisfies conversation calibration. This means that when we condition on the subsequence at which (for example), the prediction at round $k-1$ was v' and the prediction at round k was v for some $|v-v'| \geq \epsilon$, on this subsequence, the label mean was actually v. As a result, the sequence of predictions at round k must be substantially more accurate than the predictions at round $k-1$.

But neither case can occur very often: every time case (1) occurs, the fraction of conversations that makes it beyond round k is reduced by a constant factor, which can occur at most $\log(1/\delta)$ many times before only a δ fraction of conversations remain. And each time case (2) occurs, the average squared error of the predictions at round k (which makes up at least a δ fraction of days) decreases by at least $\approx \epsilon^2$ — but as the labels and predictions are both bounded in $[0,1]$, this cannot occur more than $1/(\epsilon^2\delta)$ many times.

In fact, we smoothly handle both kinds of events without explicitly breaking the analysis down into two cases, and as a result do not have to pay for the $\log(1/\delta)$ term. The following lemma is the work-horse of our analysis. It states that at any round, if the human is (perfectly) conversation calibrated given some bucketing of the model's predictions, then the squared error of the human's predictions is lower than the squared error of the model's most recent predictions by an amount scaling with ϵ^2 times the number of days that did not lead to agreement at that round — minus an error term that depends on the coarseness of the bucketing function g_h defining the human's conversation calibration guarantee. A symmetric guarantee holds for the model.

Lemma 4.3. *If the human is $(0, g_h(T))$-conversation-calibrated, then for any even k,*

$$\mathrm{SQE}(p_h^{1:T,k}, y^{T^{\geq k}}) \leq \mathrm{SQE}(p_m^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_h(T))^2|T^{\geq k+1}| + g_h(T)T \tag{1}$$

And if the model is $(0, g_h(T))$-conversation-calibrated, for any odd k,

$$\mathrm{SQE}(p_m^{1:T,k}, y^{T^{\geq k}}) \leq \mathrm{SQE}(p_h^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_m(T))^2|T^{\geq k+1}| + g_m(T)T \tag{2}$$

Proof. Let $T_k^{i,p_h} = \{t : p_h^{t,k} = p_h \textit{ and } p_m^{t,k-1} \in B_i(\frac{1}{g(T)})\}$ be the subsequence of days such that the human predicts p_h in round k and the model predicts in bucket $B_i(\frac{1}{g(T)})$ in round $k-1$. Let $m_k^{i,p_h} = \frac{\sum_{t \in T_k^{i,p_h}} y^t}{|T_k^{i,p_h}|}$ be the true mean on this subsequence. The difference in squared errors can be written as

$$\sum_{t\in T_k^{i,p_h}} (p_m^{t,k-1} - y^t)^2 - \sum_{t\in T_k^{i,p_h}} (p_h^{t,k} - y^t)^2$$

$$= \left[\sum_{t\in T_k^{i,p_h}} (p_m^{t,k-1} - y^t)^2 - \sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2\right] - \left[\sum_{t\in T_k^{i,p_h}} (p_h^{t,k} - y^t)^2 - \sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2\right]$$

(Adding and subtracting $\sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2$)

$$\geq \left[\sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - y^t)^2 - |T_k^{i,p_h}|\cdot g_h(T) - \sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2\right] -$$

$$\left[\sum_{t\in T_k^{i,p_h}} (p_h^{t,k} - y^t)^2 - \sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2\right]$$

(By Lemma A.2)

$$= \left[\sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - m_k^{i,p_h})^2 - |T_k^{i,p_h}|\cdot g_h(T)\right] - \left[\sum_{t\in T_k^{i,p_h}} (p_h^{t,k} - y^t)^2 - \sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2\right]$$

(By Lemma A.1)

$$= \left[\sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - m_k^{i,p_h})^2 - |T_k^{i,p_h}|\cdot g_h(T)\right] - \left[\sum_{t\in T_k^{i,p_h}} (p_h - y^t)^2 - \sum_{t\in T_k^{i,p_h}} (m_k^{i,p_h} - y^t)^2\right]$$

(As by definition of T_k^{i,p_h}, $p_h^{t,k} = p_h$)

$$\geq \left[\sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - m_k^{i,p_h})^2 - |T_k^{i,p_h}|\cdot g_h(T)\right] - \left[\sum_{t\in T_k^{i,p_h}} (p_h - m_k^{i,p_h})^2\right]$$

(By Lemma A.1)

$$\geq -|T_k^{i,p_h}|\cdot g_h(T) + \sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - p_h)^2$$

(As the human is $(0, g_h(T))$-conversation calibrated, $p_h = m_k^{i,p_h}$)

Summing this up for all i, p_h:

$$\sum_{\forall i,p_h} \left(-|T_k^{i,p_h}|\cdot g_h(T) + \sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - p_h)^2\right)$$

$$\geq -g_h(T)T + \sum_{\forall i,p_h} \sum_{t\in T_k^{i,p_h}} (i\cdot g_h(T) - p_h)^2$$

(As $g_h(T)$ is independent of i and p_h, and $\sum_{\forall i,p_h} \left|T_k^{i,p_h}\right| \leq T$)

$$\geq -g_h(T)T + \sum_{\forall i,p_h} \sum_{t\in T_k^{i,p_h}} \mathbb{1}[|i\cdot g_h(T) - p_h^{t,k}| \geq \epsilon - g_h(T)](i\cdot g_h(T) - p_h)^2$$

$$\geq -g_h(T)T + (\epsilon - g_h(T))^2 \sum_{\forall i,p_h} \sum_{t\in T_k^{i,p_h}} \mathbb{1}[|i\cdot g_h(T) - p_h^{t,k}| \geq \epsilon - g_h(T)]$$

Note that, for all days in the subsequence T_k^{i,p_h}, in round $k-1$ the model predicted in bucket $B_i(\frac{1}{g_h(T)}) = i\cdot g_h(T)$, and therefore in each of these days, by the definition of our bucketing, $p_m^{t,k-1} \geq (i-1)\cdot g_h(T)$ and $p_m^{t,k-1} \leq i\cdot g_h(T)$. So consider any round $t \in T_k^{i,p_h}$. If $|p_h^{t,k} - p_m^{t,k-1}| \geq \epsilon$, then we have:

$$\begin{aligned}
|p_h^{t,k} - p_m^{t,k-1}| &\leq |p_h^{t,k} - i \cdot g_h(T)| + |i \cdot g_h(T) - p_m^{t,k-1}| \\
&= |p_h^{t,k} - i \cdot g_h(T)| + i \cdot g_h(T) - p_m^{t,k-1} \\
&\leq |p_h^{t,k} - i \cdot g_h(T)| + i \cdot g_h(T) - (i-1) \cdot g_h(T) \\
&= |p_h^{t,k} - i \cdot g_h(T)| + g_h(T), \\
\implies |p_h^{t,k} - i \cdot g_h(T)| &\geq |p_h^{t,k} - p_m^{t,k-1}| - g_h(T) \geq \epsilon - g_h(T).
\end{aligned}$$

Thus, if $|p_h^{t,k} - p_m^{t,k-1}| \geq \epsilon$, then $|i \cdot g_h(T) - p_h^{t,k}| \geq \epsilon - g_h(T)$, $\forall t \in T_k^{i,p_h}$. Therefore the set of days for which the former condition holds is a subset of the latter condition, and we can write

$$\begin{aligned}
&- g_h(T)T + (\epsilon - g_h(T))^2 \sum_{\forall i, p_h} \mathbb{1}[|i \cdot g_h(T) - p_h| \geq \epsilon - g_h(T)] \cdot \left|T_k^{i,p_h}\right| \\
&\geq -g_h(T)T + (\epsilon - g_h(T))^2 \sum_{\forall i, p_h} \sum_{t \in T_k^{i,p_h}} \mathbb{1}[|p_h^{t,k} - p_m^{t,k-1}| \geq \epsilon] \\
&= -g_h(T)T + (\epsilon - g_h(T))^2 |T^{\geq k+1}|
\end{aligned}$$

(As on every day where there is a next round, the human and the model disagreed by at least ϵ)

As the human and the model are perfectly symmetrical, we also obtain the symmetrical result for the model. □

Next, we extend Lemma 4.3 to the case in which the conversation calibration error is not 0, but rather controlled by some function $f_h(\cdot)$. The idea is straightforward. We know from Lemma 4.3 that squared error would decrease significantly if the human's predictions were perfectly conversation calibrated. In fact, all we know is that the human's predictions are *close* (in ℓ_1 distance) to perfectly conversation calibrated predictions. But this is good enough, because squared error is Lipschitz, and so small changes in predictions result in small changes in squared error. As a result, approximate conversation calibration is also enough to let us bound the decrease in error across adjacent rounds:

Theorem 4.4. *If the Human is $(f_h(\cdot), g_h(\cdot))$-conversation-calibrated, then after engaging in the iterated calibration protocol for T days:*

$$\text{SQE}(p_h^{1:T,k}, y^{T^{\geq k}}) \leq \text{SQE}(p_m^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_h(T))^2 |T^{\geq k+1}| + g_h(T)T + 3\frac{f_h(g_h(T) \cdot T)}{g_h(T)} \tag{3}$$

And if the Model is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated, then after engaging in the iterated calibration protocol for T days:

$$\text{SQE}(p_m^{1:T,k}, y^{T^{\geq k}}) \leq \text{SQE}(p_h^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_m(T))^2 |T^{\geq k+1}| + g_m(T)T + 3\frac{f_m(g_m(T) \cdot T)}{g_m(T)} \tag{4}$$

Proof. Let $T_m(k,i) = \{t : p_m^{t,k-1} \in B_i\left(\frac{1}{g_h(T)}\right)\}$ be the subsequence of days in which the models predicts in bucket $B_i(\frac{1}{g_h(T)})$ at round $k-1$.

Note that the human has distance to calibration of $f_h(|T_m(k,i)|)$ on every such subsequence defined this way. Therefore, for predictions $p_h^{1:T,k}$ from the human at round k:

$$
\begin{aligned}
\mathrm{CalDist}(p_h^{1:T,k}, y^{T^{\geq k}}) &= \min_{q^{1:T} \in C(y^{1:T})} \|p_h^{1:T,k} - q^{1:T,k}\|_1 \\
&\leq \sum_{i=1}^{\frac{1}{g_h(T)}} \min_{q^{1:|T_m(k,i)|} \in C^{T_m(k,i)}(y^{1:T,k})} \|p^{1:T} - q_v^{1:T}\|_1 \\
&\leq \sum_{i=1}^{\frac{1}{g_h(T)}} f_h(|T_m(k,i)|) && \text{(By the calibration distance of the Human)} \\
&\leq \frac{f_h(g_h(T) \cdot |T^{\geq k}|)}{g_h(T)} && \leq \frac{f_h(g_h(T) \cdot T)}{g_h(T)} \\
& && \text{(By the assumption that } f_h \text{ is concave)}
\end{aligned}
$$

Let $q^{T^{\geq k}}$ be the set of perfectly calibrated predictions that are $f_h(|T_m(k,i)|)$-close to $p_h^{1:T,k}$. Then, by Lemma A.3,

$$
\begin{aligned}
\mathrm{SQErr}(p_h^{T^{\geq k},k}, y^{T^{\geq k}}) &\leq \mathrm{SQErr}(q^{T^{\geq k}}, y^{T^{\geq k}}) + 3\frac{f_h(g_h(T) \cdot T)}{g_h(T)} \\
&\leq \mathrm{SQErr}(p_m^{T_k,k-1}, y^{T^{\geq k}}) - (\epsilon - g_h(T))^2 |T^{\geq k+1}| + g_h(T)T + 3\frac{f_h(g_h(T) \cdot T)}{g_h(T)} \\
& \text{(By Lemma 4.3)}
\end{aligned}
$$

As the Human and the Model are symmetric, we also obtain the symmetric result for the Model. □

Proof of Theorem 4.1. By composing the two results in Theorem 4.4, until k such that $|T^{\geq k}| \leq \delta \cdot T$, we see that

$$
\begin{aligned}
&\mathrm{SQErr}(p_h^{T^{\geq k},k-2}, y^{T^{\geq k}}) - \mathrm{SQErr}(p_h^{1:T,k}, y^{T^{\geq k}}) \\
&\geq (\epsilon - g_h(T))^2 |T^{\geq k+1}| + (\epsilon - g_m(T))^2 |T^{\geq k}| - g_m(T)T - 3\frac{f_m(g_m(T) \cdot T)}{g_m(T)} - g_h(T)T - 3\frac{f_h(g_h(T) \cdot T)}{g_h(T)} \\
&\geq \left((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2\right) |T^{\geq k+1}| - (g_m(T) + g_h(T))T - 3\left(\frac{f_m(g_m(T) \cdot T)}{g_m(T)} + \frac{f_h(g_h(T) \cdot T)}{g_h(T)}\right)
\end{aligned}
$$

Thus, consider any round r such that $|T^{\geq r}| \geq \delta T$. By applying this expression recursively, we can bound the squared error of the model at round r by

$$
\begin{aligned}
&\mathrm{SQErr}(p_h^{1:T,r}, y^{1:T,k}) \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - ((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2)\left(\sum_{k=1}^{r} |T^{\ge k}|\right) + (g_m(T) + g_h(T))\left(\sum_{k=1}^{r} |T^{\ge k}|\right) \\
&\qquad + 3\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)\left(\sum_{k=1}^{r} 1\right) \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - ((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2)\left(\sum_{k=1}^{r} |T^{\ge k}|\right) + (g_m(T) + g_h(T))(r)T \\
&\qquad + 3\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)(r) && \text{(As } |T^{\ge k}| \le T\text{)} \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - ((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2)(r)\delta T + 2(g_m(T) + g_h(T))(r)T \\
&\qquad + 3\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)(r) \\
&\qquad\qquad\qquad\qquad\qquad\qquad && \text{(As for all } T^{\ge k} \text{ such that } k \le r\text{, } |T^{\ge k}| \ge \delta T\text{)} \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - r\epsilon^2\delta T + 3r(g_m(T) + g_h(T))T + 3r\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right) \\
&= \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - r\left(\epsilon^2\delta T + T\beta(T)\right)
\end{aligned}
$$

This completes the second part of the Theorem.

By definition, the squared error is non-negative. Therefore, we have that

$$
\begin{aligned}
&\mathrm{SQErr}(p_h^{1:T,r}, y^{1:T,k}) \le \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - r\left(\epsilon^2\delta T + T\beta(T)\right) \\
\implies &0 \le \mathrm{SQErr}(p_h^{T^{\ge r},2}, y^{1:T,k}) - r\left(\epsilon^2\delta T + T\beta(T)\right) \\
\implies &0 \le T - r\left(\epsilon^2\delta T + T\beta(T)\right) && \text{(As the maximum squared error is } T\text{)} \\
\implies &r \le \frac{1}{\epsilon^2\delta + \beta(T)}
\end{aligned}
$$

This completes the first part of the Theorem. □

We now interrogate what conversation-calibration rates are sufficient to get fast convergence to the agreement bounds we quoted in Corollary 4.2.

Theorem 4.5. *Fix any $0 < \alpha < 1$. There exists a constant γ such that if the Human is $(f_h(\cdot), g_h(\cdot))$-conversation calibrated and the model is $(f_m(\cdot), g_m(\cdot))$-conversation calibrated such that:*

$$
f_h(\tau), f_m(\tau) \in O(\tau^{\alpha}) \text{ and } g_m(\tau), g_h(\tau) \in O(\tau^{\gamma})
$$

then for every $T \ge \Omega\left((\frac{1}{\delta\epsilon^2})^{\frac{2-\alpha}{1-\alpha}}\right)$, if the agreement protocol is run for at least T days, then on a $1-\delta$ fraction of days, the two parties reach ϵ-agreement after at most $O\left(\frac{1}{\delta\epsilon^2}\right)$ rounds.

Proof. By Corollary 4.2, if $\beta(T) \le \frac{\delta\epsilon^2}{2}$, the number of rounds until agreement is at most $O(\frac{1}{\delta\epsilon^2})$. Thus we will find a sufficiently large value of T to ensure this.

We have that

$$
\beta(T) = 3(2T^{\gamma} + 2T^{\alpha(\gamma+1)-\gamma-1})
$$

This is minimized by solving for

$$
\gamma = \alpha(\gamma+1) - \gamma - 1 \implies 2\gamma - \alpha\gamma = \alpha - 1 \implies \gamma = \frac{\alpha-1}{2-\alpha}
$$

Thus we get $\beta = 12T^{\frac{\alpha-1}{2-\alpha}}$. Now, in order to ensure that $\beta(T) \leq \frac{\delta\epsilon^2}{2}$, we need

$$12T^{\frac{\alpha-1}{2-\alpha}} \leq \frac{\delta\epsilon^2}{2} \implies T^{\frac{1-\alpha}{2-\alpha}} \geq \frac{24}{\delta\epsilon^2} \implies T \geq \left(\frac{24}{\delta\epsilon^2}\right)^{\frac{2-\alpha}{1-\alpha}} = O\left(\frac{1}{\delta\epsilon^2}\right)^{\frac{2-\alpha}{1-\alpha}}.$$

□

Finally, we turn to the algorithmic problem. There are existing simple, efficient algorithms that can make sequential predictions in adversarial environments that guarantee diminishing distance to calibration at favorable rates Arunachaleswaran et al. [2025]. What we give here is an efficient reduction that takes as input an arbitrary initial model, and by reduction to a sequential prediction algorithm that achieves distance to calibration at some rate, outputs an algorithm for the model that can interact in Protocol 3.1 and against any sequence of predictions for the human, guarantee conversation calibration at the same rate. The reduction is straightforward: We use the initial model to make the round 1 predictions, and then intialize a collection of distance-to-calibration algorithms, for each round k and for each possible bucketing of the other agent's predictions. Then, at each round k, we predict according to the instance of the distance-to-calibration algorithm corresponding to that round and the bucketing of the other agent's most recent prediction.

Protocol 1 Almost-One-Step-Ahead (AOSA) Arunachaleswaran et al. [2025]

```
Input Sequence of outcomes $y^{1:T} \in \{0,1\}^T$
Output Sequence of predictions $p^{1:T} \in \{0, \frac{1}{m}, ..., 1\}^T$ for $m = 1/\sqrt{T}$
for $t = 1 \ldots, T$ do
    Given look-ahead predictions $\tilde{p}^{1:t-1}$, define the look-ahead bias conditional on a prediction $p$ as:
        $\alpha_{\tilde{p}^{1:t-1}}(p) := \sum_{s=1}^{t-1} \mathbb{I}[\tilde{p}^s = p](\tilde{p}^s - y^s)$
    Choose two adjacent points $p_i = \frac{i}{m}, p_{i+1} = \frac{i+1}{m}$ satisfying:
        $\alpha_{\tilde{p}^{1:t-1}}(p_i) \leq 0$ and $\alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \geq 0$
    Arbitrarily predict $p^t = p_i$ or $p^t = p_{i+1}$    Upon observing the (adversarially chosen) outcome $y^t$, set look-ahead prediction
        $\tilde{p}^t = \arg\min_{p \in \{p_i, p_{i+1}\}} |p - y^t|$
```

First we quote the distance to calibration guarantee of Almost-One-Step-Ahead (Algorithm 1)

Theorem 4.6 (Arunachaleswaran et al. [2025])**.** *Algorithm 1 (Almost-One-Step-Ahead) guarantees that against any sequence of outcomes,* $\mathrm{CalDist}(p^{1:T}, y^{1:T}) \leq 2\sqrt{T} + 1$.

As stated, Algorithm 1 is defined as a function of the length T of the sequence on which it will be evaluated. In our reduction, we will want guarantees that hold over many different sequences whose lengths we do not know ahead of time. However, it is not hard to convert Algorithm 1 into an algorithm that has similar bounds and does not require knowing T ahead of time, using a doubling trick. We give such an algorithm in Algorithm 4 in Appendix A.

Theorem 4.7. *Algorithm 4 (Almost-One-Step-Ahead with unknown T) guarantees that against any sequence of outcomes,* $\mathrm{CalDist}(p^{1:T}, y^{1:T}) \leq O(\sqrt{T})$.

The proof is in Appendix A.

Finally, we can give our reduction (Algorithm 2) that takes as input an initial model $M_0 : \mathcal{X}_m \to \mathcal{Y}$, a sequential prediction algorithm D with a concave bound $f_m(\cdot)$ on its distance to calibration, and a bucketing function $g_m(\cdot)$. We show that for any discretization function g_m, Algorithm 2 guarantees $(f_m(\cdot), g_m(\cdot))$ conversation calibration against any sequence of outcomes and predictions of the human. We will refer to the prediction made by Algorithm 2 in round 1 (for any day t) as $\textsc{Converse}_1$.

Protocol 2 CONVERSE$(M_0, D, g_m(T))$: A reduction from an online decision-making algorithm to an algorithm with low conversation-calibration error

```
Input Baseline model algorithm $M_0$, D2C algorithm $D$, Discretization $g_m(T)$
We denote $D_{k,i}$ as an instantiation of $D$ which is given as input only the subsequence of days where $p_h^{t,k} \in [(i-1) \cdot g_m(T), i \cdot g_m(T)]$, and denote $D_{k,i,t}$ be the prediction of $D_{k,i}$ at round $t$.
for $t = 1, \ldots, T$ do
    Receive $x_m^t$
    Send prediction $p_m^{t,1} = M_0(x_m^t)$ to human
    for $k = 3, 5, \ldots$ do
        Initialize empty set $S$
        Observe human prediction $p_h^{t,k-1}$
        if $|p_h^{t,k-1} - p_m^{t,k-2}| < \epsilon$ then
            Predict $p_h^{t,k-1}$ and break out of loop
        Let $i$ be such that $p_h^{t,k-1} \in [(i-1) \cdot g_m(T), i \cdot g_m(T)]$
        if $D_{k-1,i}$ uninitialized then
            Initialize $D_{k-1,i}$
        Send prediction $p_m^{t,k} = D_{k-1,i,t}$ to human
        $S \leftarrow S \cup (k-1, i)$
        if $|p_h^{t,k-1} - p_m^{t,k}| < \epsilon$ then
            Predict $p_h^{t,k-1}$ and break out of loop
    Observe $y^t$
    for $(k, i) \in S$ do
        Update $D_{k,i}$ with $(D_{k,i,t}, y^t)$
```

Theorem 4.8. *If D has worst-case* CalDist *of $f_m(\cdot)$, then for any bucketing function $g_m(\cdot)$,* CONVERSE$(M_0, D, g_m(\cdot))$ *is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated in the worst case over label outcomes and conversations. Moreover the first round prediction of* CONVERSE *are the same as the prediction of the base model M_0 for all t:* CONVERSE$_1(x_m^t) = M_0(x_m^t)$*, for all t.*

Proof. We first must show that CONVERSE$(M_0, D, g_m(\cdot))$ is $f_m(\cdot), g_m(\cdot)$-conversation-calibrated. Observe that Algorithm 2 instantiates a new copy of the algorithm D for each round k, bucket i pair corresponding to each of the sets $T_m(k, i)$. This immediately implies that the sequence of predictions made by each instance $D_{k,i}$ satisfies

$$\text{CalDist}(p_h^{T_m(k,i),k}, y^{T_m(k,i)}) \leq f(|T_m(k,i)|),$$

by assumption that D has worst-case distance to calibration of $f_m(\cdot)$. Since this holds for the sequence of predictions for all round k, bucket i pairs, by its corresponding instance of $D_{k,i}$, we have that CONVERSE$(M_0, D, g_m(\cdot))$ is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated. That CONVERSE$_1(x_m^{1:T}) = M_0(x_m^{1:T})$, for all t follows by construction. □

Corollary 4.9. CONVERSE$(M_0, \text{AOSA}, T^{\frac{-1}{3}})$ *is $(\sqrt{T}, T^{\frac{-1}{3}})-$conversation-calibrated, and* CONVERSE$_1(x_m^t) = M_0(x_m^t)$*, for all t.*

Finally, we end with a corollary putting all of our results together. We can take any initial model M_0 and efficiently convert it into a protocol that can engage in conversations with a human. If the human satisfies conversation calibration (a significantly weaker assumption than Bayesian rationality), then not only will the conversations halt quickly, but they will result in outcomes that are only more accurate than either the human or the model's initial judgments. This holds despite the fact that we have no assumptions on the form of the initial model M_0, so it can be the result of an arbitrarily sophisticated machine learning process. We state the corollary as if both parties are implemented using CONVERSE to be concrete about computational tractability and so that we can be specific about rates, but this is not necessary — the only important thing is that both parties are conversation calibrated.

Corollary 4.10. *If the human runs* CONVERSE$(M_0^h, \text{AOSA}, g_h(\cdot))$ *and the model runs* CONVERSE$(M_0^m, \text{AOSA}, g_m(\cdot))$ *for $g_m(T) = g_h(T) = T^{-\frac{1}{3}}$, then:*

- *For any $\epsilon, \delta \in [0,1]$, on a $1-\delta$ fraction of days, they reach ϵ-agreement after at most K rounds of conversation where: $K \leq \frac{1}{\epsilon^2\delta - 6T^{-\frac{1}{3}}}$*
- *For the subsequence of days that make it to round k s.t. $|T^{\geq k}| \geq \delta T$ with associated human prediction subsequences $p_h^{T^{\geq k},k}$ and outcome subsequences $y^{T^{\geq k}}$, we have that for $M_0 \in \{M_0^m, M_0^h\}$: $\frac{\mathrm{SQErr}(p_h^{T^{\geq k},k}, y^{T^{\geq k}})}{T} \leq \frac{\mathrm{SQErr}(M_0, y^{T^{\geq k}})}{T} - i(\epsilon^2\delta - \frac{12}{T^{\frac{1}{3}}})$.*

5 Agreement in d Dimensions

We now extend our results from 1 dimensional label spaces to d dimensional label spaces: $\mathcal{Y} = [0,1]^d$ (the setting given in Definition 3.6). As in the previous section, here predictions are the same as messages ($p_m^{k,t} = \hat{y}_m^{k,t}$ and $p_h^{k,t} = \hat{y}_h^{k,t}$), and therefore in this section we will refer to both these terms as $p_m^{k,t}$ (and $p_h^{k,t}$) for simplicity.

At a high level, our argument will be similar. We will measure the error of our predictions using the sum of the squared error in each of the coordinates of our predictions, which will also serve as our potential function. We will continue to argue that at each round, either many conversations end, or else the squared error of the predictions must substantially improve, limiting the number of rounds of conversation that can occur. Of course, the maximum squared error is now d, rather than 1, and so the number of rounds until agreement will be larger by a factor of d. There is another step in the argument within which one must be careful not to lose another factor of d. For tractability, we have only asked for conversation calibration conditions to hold *marginally* on each coordinate of our predictions. So, we need to argue that error decreases coordinate-wise. But imagine a sequence of predictions at round k on which we have not reached ϵ-agreement. It might be that each prediction agrees with the previous round's predictions on all but a single coordinate — and hence it might be that for any *particular* coordinate, there is in fact ϵ-disagreement with the prior round's prediction *in that coordinate* on only a $1/d$ fraction of the rounds. We are able to avoid losing another factor of d in our analysis by keeping more careful track of the error in each coordinate — since if the disagreements are uniformly spread across all of the coordinates, although it is true that we do not improve the error by as much in each coordinate, we are able to improve the error in all coordinates simultaneously.

In our analysis, we will often need to focus on a single coordinate j of the multi-dimensional prediction or label: we will write $p_h^{t,k}[j]$ or $y[j]$ to denote the value at this coordinate. We measure accuracy using the following multi-dimensional extension of our squared error definition:

Definition 5.1 (Multi-Dimension Squared Error)**.** *The squared error of a sequence of $d-$dimensional predictions p with respect to the $d-$dimensional outcomes y is:*

$$\mathrm{SQE}(p^{1:T}, y^{1:T}) = \sum_{j \in [d]} \mathrm{SQE}(p^{1:T}[j], y^{1:T}[j]).$$

Theorem 5.2. *If the Human is (f_h, g_h)-conversation-calibrated and the Model is (f_m, g_m)-conversation-calibrated in d dimensions, then for any $\epsilon, \delta \in [0,1]$, on a $1-\delta$ fraction of days, they reach ϵ-agreement after at most K rounds of conversation for*

$$K \leq \frac{d}{\epsilon^2\delta - \beta(T)}$$

where $\beta(T) = 3d\left(g_m(T) + g_h(T) + \frac{f_m(g_m(T)\cdot T)}{g_m(T)\cdot T} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)\cdot T}\right)$, a term that will tend to 0 for appropriately instantiated functions g and f.

Furthermore, for any round k such that $|T^{\geq k}| \geq \delta T$, we have that

$$\frac{\mathrm{SQErr}(p_h^{T^{\geq k},k}, y_h^{T^{\geq k},k})}{T} \leq \frac{\mathrm{SQErr}(p_h^{T^{\geq k},2}, y_h^{T^{\geq k},k})}{T} - k(\epsilon^2\delta - \beta(T)).$$

In other words, each round of conversation is error improving compared to the initial predictions of the human (or the model), with the error improving at a rate that is linear in the number of rounds of conversation.

A corollary of this theorem is that after T is taken to be sufficiently large, agreement occurs rapidly on almost every day, and for $(1-\delta)$ of the days, each further round of conversation leads to an $\epsilon^2\delta$ decrease in squared error.

Corollary 5.3. *When $\beta(T) \leq \frac{\delta\epsilon^2}{2}$, on a $1-\delta$ fraction of days, the number of rounds until agreement is at most $K \leq \frac{2d}{\delta\epsilon^2}$.*

Finally, in Corollary 5.7 we give a reduction that allows us to convert an arbitrary model into an algorithm that satisfies $(\sqrt{T}, T^{\frac{-1}{3}})$-conversation calibration in the d-dimensional setting.

We can now state our main work-horse lemma, which again holds for *perfectly* conversation calibrated predictions. It states that at any round k, the squared error of the vector-valued predictions must decrease compared to the squared error at the previous round, in proportion to ϵ^2 and the fraction of days that do not lead to agreement at round k. We then extend the argument to predictions that have conversation-calibration error that is controlled by some function $f_h(\cdot)$. Since the arguments follow a similar analysis to those in Section 4, applied once to each dimension d, we defer all proofs to Appendix B.

Lemma 5.4. *If the human is $(0, g_h(T))$-conversation-calibrated for d-dimensional vector predictions, then for any even k,*

$$\mathrm{SQE}(p_h^{1:T,k}, y^{T^{\geq k}}) \leq \mathrm{SQE}(p_m^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_h(T))^2|T^{\geq k+1}| + dg_h(T)T$$

And if the model is $(0, g_h(T))$-conversation-calibrated, for any odd k,

$$\mathrm{SQE}(p_m^{1:T,k}, y^{T^{\geq k}}) \leq \mathrm{SQE}(p_h^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_m(T))^2|T^{\geq k+1}| + dg_m(T)T \tag{5}$$

Theorem 5.5. *If the Human is $(f_h(\cdot), g_h(\cdot))$-conversation-calibrated in d dimensions, then after engaging in the iterated calibration protocol for T days:*

$$\mathrm{SQE}(p_h^{1:T,k}, y^{T^{\geq k}}) \leq \mathrm{SQE}(p_h^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_h(T))^2|T^{\geq k+1}| + d\cdot g_h(T)T + 3d\cdot\frac{f_h(g_h(T)\cdot T)}{g_h(T)}$$

And if the Model is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated in d dimensions, then after engaging in the iterated calibration protocol for T days:

$$\mathrm{SQE}(p_m^{1:T,k}, y^{T^{\geq k}}) \leq \mathrm{SQE}(p_m^{T^{\geq k},k-1}, y^{T^{\geq k}}) - (\epsilon - g_m(T))^2|T^{\geq k+1}| + d\cdot g_m(T)T + 3d\cdot\frac{f_m(g_m(T)\cdot T)}{g_m(T)}.$$

Now, similarly to Section 4, we introduce our reduction (Algorithm 5) that takes as input an initial model $M_0 : \mathcal{X}_m \to \mathcal{Y}$, a sequential prediction algorithm D with a concave bound $f_m(\cdot)$ on its distance to calibration, and a bucketing function $g_m(\cdot)$. We show that Algorithm 5 efficiently guarantees $(f_m(\cdot), g_m(\cdot))$-conversation calibration against any sequence of outcomes and predictions of the human.

Theorem 5.6. *If D has worst-case* CalDist *of $f_m(\cdot)$, then for any bucketing function $g_m(\cdot)$,* CONVERSE-DDIM*$(M_0, D, g_m(\cdot))$ is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated, and for any sequence of labels $y^{1:T}$, the first round prediction of* CONVERSE-DDIM *is the same as the prediction of the base model M_0 for all t:* CONVERSE-DDIM$_1(x_m^t) = M_0(x_m^t)$*, for all t.*

Corollary 5.7. *Algorithm 5* CONVERSE-DDIM*$(M_0, \mathrm{AOSA}, T^{\frac{-1}{3}})$ is $(\sqrt{T}, T^{\frac{-1}{3}})$-conversation-calibrated.*

We conclude with a final corollary putting the above together. Any arbitrary baseline model can be efficiently converted into a protocol that interacts with a human, and if this human satisfies our conversation-calibration condition, conversations will reach agreement quickly.

Corollary 5.8. *If the human runs* CONVERSE-DDIM$(M_0^h, \mathrm{AOSA}, g_h(\cdot))$ *and the model runs* CONVERSE-DDIM$(M_0^m, \mathrm{AOSA}, g_m(\cdot))$ *for $g_m(T) = g_h(T) = T^{-\frac{1}{3}}$, then:*

- *For any $\epsilon, \delta \in [0,1]$, on a $1-\delta$ fraction of days, they reach ϵ-agreement after at most K rounds of conversation where: $K \leq \frac{d}{\epsilon^2\delta - T^{-\frac{1}{3}}}$.*

- *For the subsequence of days that make it to round k s.t. $|T^{\geq k}| \geq \delta T$ with associated human prediction subsequences $p_h^{T^{\geq k},k}$ and states of nature subsequences $y^{T^{\geq k}}$, we have that $\frac{\mathrm{SQErr}(p^{T^{\geq k},k}, y^{T^{\geq k}})}{T} \leq \frac{\mathrm{SQErr}(M_0, y^{T^{\geq k}})}{T} - i(\epsilon^2\delta - \frac{12}{T^{\frac{1}{3}}})$.*

6 Agreement when Communicating Decisions

We now turn our attention to the action feedback setting (Setting 3.8). Recall that in this setting, the label space $\mathcal{Y} \subseteq [0,1]^d$ is high dimensional, and the parties communicate with one another not by providing point predictions $\hat{y} \in \mathcal{Y}$, but rather by communicating the action a in an action space $\mathcal{A}$ that is utility maximizing according to their predictions. In this section, the messages $p_m^{t,k}$ and $p_h^{t,k}$ denote the actions which the Human and Model communicate at each round. Note that by definition of Setting 3.8, $p_m^{t,k}$ is the optimal action given the Model's prediction of the label vector $\hat{y}_m^{t,k}$ (and the equivalent statement holds for the Human).

Rather than arguing that the squared error of the predictions decreases at each round of conversation, we will argue that the *utility* of the sequence of communicated actions will increase at each iteration. Towards this end we will define shorthand notation that expresses the *summed* utility of a sequence of actions over time, with respect to a sequence of outcomes.

Definition 6.1. *Fix any utility function U as defined in Definition 1. We extend our notation to allow U to take as input a* sequence *of communicated actions $p^{1:T}$ and a corresponding sequence of outcomes $y^{1:T}$ by letting this denote the* summed *utility as computed over this sequence:*

$$U(p^{1:T}, y^{1:T}) = \sum_{t=1}^{T} U(p^t, y^t)$$

We will prove the following theorem in this section:

Theorem 6.2. *If the Human is $f_h(\cdot)$-decision-conversation-calibrated and the Model is $f_m(\cdot)$-decision-conversation-calibrated, then on a $1-\delta$ fraction of days, they reach ϵ-agreement in at most*

$$K \leq \frac{1}{2\epsilon\delta - \gamma(T)} + 1$$

rounds, where $\gamma(T) = \frac{2Ld|A|^2 \cdot f_h(\frac{T}{|A|^2}) + 2Ld|A|^2 \cdot f_m(\frac{T}{|A|^2})}{T}$ is a term that will tend to 0 as T grows large. Furthermore, for any round k such that $|T^{\geq k}| \geq \delta T$,

$$U(p_h^{1:T,k}, y^{1:T}) \geq U(p_m^{T^{\geq k},k-1}, y^{1:T}) + kT(2\epsilon\delta - \gamma(T))$$

Corollary 6.3. *When $\gamma(T) \leq \varepsilon\delta$, on a $1-\delta$ fraction of days, the number of rounds until agreement is at most*

$$K \leq \frac{1}{\varepsilon\delta} + 1.$$

And for any round k such that $|T^{\geq k}| \geq \delta T$,

$$U(p_h^{1:T,k}, y^{1:T}) \geq U(p_m^{T^{\geq k},k-1}, y^{1:T}) + k\epsilon\delta T$$

The proof follows a similar structure to the proof of our agreement theorem in the canonical setting (Theorem 4.1). At a high level, we analyze each *round k* of communication separately, across days. Intuitively there are again two cases. In the first case, most of the conversations that make it to round k end in agreement. Again, this is a good case, as we wish to show that most conversations end in agreement quickly. In the remaining case, most of the predictions made at round k ϵ-disagree. Here our argument differs: Since the parties are not communicating their (d-dimensional) predictions directly, we cannot argue that the squared error of the predictions at round k decreases. However, our notion of conversation-decision-calibration does allow us to argue that the average *utility* of the predictions made at round k increases substantially compared to the prior round. Thus the downstream utility of the human takes the role of squared error in our potential argument (and is what allows us to argue that the conversations are utility increasing). In fact, because our utility functions are linear, compared to the canonical setting, this allows us to get an improved rate of convergence — depending now on $1/\epsilon$ rather than $1/\epsilon^2$. The below lemma formalizes the progress that we make at round k of a conversation, across days:

Lemma 6.4. *If the Human is $f_h(\cdot)$-decision-conversation-calibrated, then after engaging in Protocol 3.1 instantiated in the action feedback setting (Definition 3.7) for T days, for all* odd *rounds k:*

$$U(p_h^{1:T,k}, y^{1:T,k}) - U(p_m^{1:T,k-1}, y^{1:T,k}) \geq \epsilon|T^{\geq k+1}| - 2Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right)$$

Furthermore, if the Model is $f_m(\cdot)$-decision-conversation-calibrated, then after engaging in Protocol 3.1 instantiated in the action feedback setting (Definition 3.7) for T days, for all even *rounds k:*

$$U(p_m^{1:T,k}, y^{T^{\geq k}}) - U(p_h^{1:T,k-1}, y^{T^{\geq k}}) \geq \epsilon|T^{\geq k+1}| - 2Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right)$$

Proof. Let $T_k^{a_h,a_m} = \{t : p_h^{t,k} = a_h \textit{ and } p_m^{t,k-1} = a_m\}$ be the subsequence of days such that the human sends the message a_h in round k and the model sends the message a_m in round $k-1$.

By definition, for all $t \in T_k^{a_h,a_m}$, $\arg\max_{a \in \mathcal{A}} U(a, p_h^{t,k}) = a_h$ and $\arg\max_{a \in \mathcal{A}} U(a, p_m^{t,k-1}) = a_m$. Then, we can write the difference in utilities as

$$
\begin{aligned}
&U(p_h^{1:T,k}, y^{T^{\geq k}}) - U(p_m^{1:T,k-1}, y^{T^{\geq k}}) \\
&= \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} U(a_h, y^t) - \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} U(a_m, y^t) \\
&= \sum_{a_h,a_m\in\mathcal{A}} U(a_h, \sum_{t\in T_k^{a_h,a_m}} y^t) - \sum_{a_h,a_m\in\mathcal{A}} U(a_m, \sum_{t\in T_k^{a_h,a_m}} y^t) \qquad \text{(By the linearity of } U(a,\cdot)\text{)} \\
&\geq \sum_{a_h,a_m\in\mathcal{A}} U\left(a_h, \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}\right) - \sum_{a_h,a_m\in\mathcal{A}} U\left(a_m, \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}\right) - \\
&2L\cdot \| \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k} - \sum_{t\in T_k^{a_h,a_m}} y^t \|_1 \qquad \text{(By the } L\text{-lipschitzness of } U(a,\cdot)\text{)} \\
&\geq \sum_{a_h,a_m\in\mathcal{A}} U\left(a_h, \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}\right) - \sum_{a_h,a_m\in\mathcal{A}} U\left(a_m, \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}\right) - 2L\cdot \sum_{j\in[d]} (\sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}[j] - \sum_{t\in T_k^{a_h,a_m}} y^t[j]) \\
&\geq \sum_{a_h,a_m\in\mathcal{A}} U\left(a_h, \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}\right) - \sum_{a_h,a_m\in\mathcal{A}} U\left(a_m, \sum_{t\in T_k^{a_h,a_m}} p_h^{t,k}\right) - 2Ld\cdot f_h(T^{h,k}_{a_m,a_h}) \\
&\qquad \text{(By the DC-calibration guarantee of the Human)} \\
&= \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} U(a_h, p_h^{t,k}) - \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} U(a_m, p_h^{t,k}) - \sum_{a_h,a_m\in\mathcal{A}} 2Ld\cdot f_h(|T_k^{a_h,a_m}|) \\
&\qquad \text{(By the linearity of } U(a,\cdot)\text{)} \\
&= \sum_{a_h,a_m\in\mathcal{A}} \left(\sum_{t\in T_k^{a_h,a_m}} (U(a_h, p_h^{t,k}) - U(a_m, p_h^{t,k}) \right) - \sum_{a_h,a_m\in\mathcal{A}} 2Ld\cdot f_h(|T_k^{a_h,a_m}|) \\
&\qquad \text{(By the linearity of } U(a,\cdot)\text{)} \\
&\geq \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} \mathbb{1}[U(a_h, p_h^{t,k}) - U(a_m, p_h^{t,k}) \geq \epsilon]\cdot\epsilon \\
&+ \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} \mathbb{1}[U(a_h, p_h^{t,k}) - U(a_m, p_h^{t,k}) < \epsilon] \left(U(a_h, p_h^{t,k}) - U(a_m, p_h^{t,k}) \right) - \sum_{a_h,a_m\in\mathcal{A}} 2Ld\cdot f_h(|T_k^{a_h,a_m}|) \\
&\geq \sum_{a_h,a_m\in\mathcal{A}} \sum_{t\in T_k^{a_h,a_m}} \mathbb{1}[U(a_h, p_h^{t,k}) - U(a_m, p_h^{t,k}) \geq \epsilon]\cdot\epsilon - \sum_{a_h,a_m\in\mathcal{A}} 2Ld\cdot f_h(|T_k^{a_h,a_m}|) \\
&\qquad \text{(As } a_h \text{ is the best response under } p_h^{t,k}, \forall t\in[T^{h,k}_{a_m,a_h}]\text{)} \\
&= \epsilon\cdot|T^{\geq k+1}| - 2Ld \sum_{a_h,a_m\in\mathcal{A}} f_h(|T_k^{a_h,a_m}|) \\
&\qquad \text{(As } T^{\geq k+1} \text{ is exactly the days on which the human at round } k \text{ does not agree with the model)} \\
&\geq \epsilon\cdot T^{\geq k+1} - 2Ld|A|^2\cdot f_h\left(\frac{|T^{\geq k}|}{|A|^2}\right) \qquad \text{(By the concavity of } f_h\text{)} \\
&\geq \epsilon\cdot|T^{\geq k+1}| - 2Ld|A|^2\cdot f_h\left(\frac{T}{|A|^2}\right)
\end{aligned}
$$

As the model and the human are symmetric, we also attain the symmetric result for the model. □

We can now prove the theorem, by iteratively applying Lemma 6.4 to each round of conversation:

Proof of Theorem 6.2. By composing both parts of Theorem 6.4, we have that, for any k,

$$U\left(p_h^{1:T,k}, y^{T^{\geq k}}\right) - U\left(p_h^{1:T,k-2}, y^{T^{\geq k}}\right) \geq \epsilon|T^{\geq k+1}| + \epsilon|T^{\geq k}| - 2Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right) - 2Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right)$$

$$\implies U\left(p_h^{1:T,k}, y^{T^{\geq k}}\right) - U\left(p_h^{1:T,2}, y^{T^{\geq k}}\right) \geq \sum_{q=1}^{k-1}\left(\epsilon|T^{\geq q+1}| + \epsilon|T^{\geq q}| - 2Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right) - 2Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right)\right)$$

(Recursively applying the result)

Let us consider any round r in which $|T^{\geq r}| \geq \delta T$. We have that:

$$\begin{aligned}
U\left(p_h^{1:T,r}, y^{1:T,r}\right) - U\left(p_h^{1:T,2}, y^{1:T,k}\right) &\geq \sum_{q=1}^{r-1}\left(\epsilon|T^{\geq q+1}| + \epsilon|T^{\geq q}| - 2Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right) - 2Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right)\right) \\
&\geq -2\left(k-1\right)Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right) - 2\left(k-1\right)Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right) + \epsilon\sum_{q=1}^{k-1}\left(|T^{\geq q+1}| + T\right) \\
&\geq -2\left(k-1\right)Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right) - 2\left(k-1\right)Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right) + \epsilon\sum_{q=1}^{k-1}\left(2\delta T\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(As } |T^{\geq q}| \text{ is} \geq \delta T\text{)} \\
&\geq \left(k-1\right)\left(2\epsilon\delta T - 2Ld|A|^2 \cdot f_h\left(\frac{T}{|A|^2}\right) - 2Ld|A|^2 \cdot f_m\left(\frac{T}{|A|^2}\right)\right) \\
&\geq \left(k-1\right)\left(2\epsilon\delta T - T\gamma\left(T\right)\right)
\end{aligned}$$

This proves the second result in the Theorem.

However, we also have that $U\left(p_h^{1:T,k}, y^{1:T,k}\right) \leq T$. Therefore, we have that

$$\begin{aligned}
&U\left(p_h^{1:T,r}, y^{1:T,r}\right) - U\left(p_h^{1:T,2}, y^{1:T,k}\right) \geq \left(k-1\right)\left(2\epsilon\delta T - T\gamma\left(T\right)\right) \\
&\implies T - U\left(p_h^{1:T,2}, y^{1:T,k}\right) \geq \left(k-1\right)\left(2\epsilon\delta T - T\gamma\left(T\right)\right) \\
&\implies T \geq \left(k-1\right)\left(2\epsilon\delta T - T\gamma(T)\right) && \text{(As } U(\cdot) \geq 0\text{)} \\
&\implies k \leq \frac{1}{2\epsilon\delta - \gamma(T)} + 1
\end{aligned}$$

This proves the first result in the Theorem. □

We now turn to the algorithmic reduction that allows us to convert a model into an algorithm capable of maintaining conversation-decision calibrated predictions. To do so, we need to define some formalism to be able to express the guarantees of the algorithm of Noarov et al. [2023], which informally, is able to maintain d dimensional predictions that are *unbiased* conditional on an arbitrary collection of specified *events*. We define a special case of these events below, which is strictly less general than the type of events supported by Noarov et al. [2023], but sufficient for our usage.

Definition 6.5 (Event indicator $E(c^t, \hat{y}^t)$)**.** *For any t, the event indicator function $E : \mathcal{C} \times \mathcal{Y} \to \{0, 1\}$ takes as input the context c^t and prediction $\hat{y}^t$ in round t, and outputs a binary indicator of whether or not event E is active.*

We write a collection of events as $\mathcal{E}$. We can now state the guarantees of the algorithm given in Noarov et al. [2023]:

Theorem 6.6 (Noarov et al. [2023])**.** *Given a convex compact d-dimensional real valued prediction space and a collection $\mathcal{E}$ of events of size $|\mathcal{E}|$, for any $0 < \alpha < 1$, the algorithm* UNBIASEDPREDICTION *outputs, for any sequence of adaptively chosen labels, a sequence of d-dimensional predictions $\hat{y}^1, \ldots, \hat{y}^T$ satisfying with probability $1 - \alpha$, for every event $E \in \mathcal{E}$ and every coordinate $j \in [d]$:*

$$\left|\sum_{t=1}^{T} E(c^t, \hat{y}^t)](\hat{y}^t[j] - y^t[j])\right| \leq O\left(\log(d\,|\mathcal{E}|\,T) + \sqrt{T\log\left(\frac{|\mathcal{E}|d}{\alpha}\right)}\right) \tag{6}$$

The per-round running time of UNBIASEDPREDICTION *is polynomial in d and $|\mathcal{E}|$.*

UnbiasedPrediction is instantiated with the set of events $\mathcal{E}$ and α and, on every day, takes as input a context pair needed to evaluate each event.

In our reduction we will run a different instantiation of the UnbiasedPrediction algorithm for each round k, and for the kth instantiation, the contexts at each day t will be the conversation $C^{t,1:k-1}$ that has taken place on that day so far.

We want our predictions at each round to be unbiased conditional on events which are defined by the human's recommended action in the previous round, and the model's recommended action in this round. We define the following event set accordingly:

Definition 6.7 (Action-Conversation Events)**.** *For each pair of actions $a_h, a_m \in \mathcal{A}$ and each round k define the event:*

$$E_{a_h,a_m,k}(\hat{y}^{t,k}, C^{t,1:k-1}) = \mathbb{1}[\arg\max_{a\in\mathcal{A}} U(a, \hat{y}^{t,k}) = a_m] \cdot \mathbb{1}[p_h^{t,k-1} = a_h]$$

Let $\mathcal{E}_k := \{E_{a_h,a_m,k} \forall a_h, a_m \in \mathcal{A}\}$.

We are now ready to define our reduction in Algorithm 3.

Protocol 3 Converse-Action(M_0, α)

```
Input Baseline model algorithm $M_0$, Discretization $g_m(T)$
for $t = 1, \dots, T$ do
  Receive $x_m^t$
  Send prediction $p_m^{t,1} = M_0(x_m^t)$ to the human
  for $k = 2, 4, 6, \dots$ do
    $L \leftarrow k$
    if $D_{k+1}$ uninitialized then
      Initialize $D_{k+1} = $ UnbiasedPrediction$(\mathcal{E}_{k+1}, \alpha)$
    Observe human action recommendation $p_h^{t,k}$
    if $p_h^{t,k} = p_m^{t,k-1}$ or $|U(p_h^{t,k}, \hat{y}_m^{t,k-1}) - U(p_m^{t,k-1}, \hat{y}_m^{t,k-1})| \leq \epsilon$ then
      Predict $p_h^{t,k}$ and break out of loop
    Set prediction $\hat{y}_m^{t,k+1} = D_{k+1}(C^{t,1:k-1})$
    Send recommendation $p_m^{t,k+1} = \arg\max_{a\in\mathcal{A}} U(a, \hat{y}_m^{t,k+1})$ to human
  Observe $y^t$
  for $k \in 2, 4, \dots, L$ do
    Update $D_{k+1}$ with $y^t$
```

Theorem 6.8. Converse-Action(M_0, α) *is* $O\left(\log(2d|\mathcal{A}|^2 T + \sqrt{T \ln\left(\frac{|\mathcal{A}|^2 d}{\alpha}\right)}\right)$*-DC-calibrated with probability* $1-\alpha$*, and for any sequence of labels* $y^{1:T}$*, its first-round prediction is the same as the prediction of the base model* M_0 *for all t:* Converse-Action$(M_0, \alpha)_1(x_m^t) = M_0(x_m^t)$*, for all t.*

Proof. By construction, in each odd round k, Converse-Action(M_0, α) runs UnbiasedPrediction$(\mathcal{E}_k, \alpha)$ with subsequences defined by $\mathcal{E}_k$ in order to obtain predictions. By Theorem 6.6, in each round, the bias on subsequences defined by the model's action recommendation and the human's action recommendation on the previous round is $O\left(\log(2d\,|\mathcal{E}|\,T) + \sqrt{T \ln\left(\frac{|\mathcal{E}|d}{\alpha}\right)}\right)$. Thus the algorithm is $O\left(\log(2d\,|\mathcal{E}|\,T) + \sqrt{T \ln\left(\frac{|\mathcal{E}|d}{\alpha}\right)}\right)$-DC-calibrated.

The second result follows directly from the definition of Converse-Action(M_0, α). □

Theorem 6.9. *Fix an L-Lipschitz utility function U. If the human runs* Converse-Action(M_0^h, α) *and the model runs* Converse-Action(M_0^m, α)*, then, if* $T \geq \frac{O\left(L^2 d^3 |A|^5 (1+\log(\frac{1}{\alpha}))\right)}{\epsilon^2 \delta^2}$*, with probability* $\geq 1 - 2\alpha$*, on a* $1-\delta$ *fraction of days, the number of rounds until agreement is at most*

$$K \leq \frac{1}{\epsilon\delta} + 1$$

Furthermore, for any round k such that $|T^{\geq k}| \geq \delta T$,

$$U(p_h^{1:T,k}, y^{1:T}) \geq U(p_m^{T^{\geq k},k-1}, y^{1:T}) + k\epsilon\delta T$$

We defer the proof to Appendix C.

7 Bayesian Agreement Theorems

In this section we show how to recover one-shot Agreement Theorems for Bayesians with a common prior, in the style of past work Aumann [1976], Geanakoplos and Polemarchakis [1982], Aaronson [2005], Kong and Schoenebeck [2023], Frongillo et al. [2023]. In most of this paper, we have studied a repeated interaction across many days, within an environment about which we have made no assumptions. Our theorems hinged on tractable calibration conditions that we imposed on the participants. In contrast, past work on agreement theorems has assumed two interlocutors who share common and complete knowledge of a *prior distribution* from which instances are drawn, and are perfect Bayesians — at each round of conversation, they condition on everything they have observed (the features of the instance they have seen, as well as the transcript of the conversation), and report their posterior expectation of the label. The strength of the approach that we have taken in most of this paper is that we do not need to assume any distributional knowledge (or even the existence of a distribution), and our assumptions on the agents are tractable (in contrast to an assumption that the agents can compute posterior distributions, which is in general intractable in large state spaces). On the other hand, our guarantees are necessarily about sequences of many interactions, whereas past work on Aumannian agreement theorems give guarantees for conversations about *single* instances, that hold with high probability over the draw of the instance from the prior distribution.

In this section, we show that our theorems are strictly more general than this one-shot setting, in that all of our theorems can be "lifted" to the one-shot setting if we are willing to make the assumption (as past work does) that instances are drawn from a commonly known prior and that the agents report correct posterior expectations. To demonstrate this, we prove two things:

1. First, we show that in the sequential setting, if the instance at each round is drawn independently from a known prior distribution, then an Agent who reports the posterior expectation of the label at each round of conversation (conditional on everything they have observed so far, including the transcript of the conversation) will satisfy our various notions of conversation calibration, no matter how their interlocutor is behaving. This result is in the spirit of Dawid [1982], and our analysis proceeds according to the following thought experiment: when arguing that the Bayesian is conversation-calibrated at some round k of the conversation, we imagine that at each day t, the label y^t is re-drawn from the Bayesian's posterior distribution on y^t at round k. This does not change the joint distribution on transcripts, and so any statement that is true of transcripts under this thought experiment is true under the original transcript distribution. But within this thought experiment, the Bayesian is always announcing the true mean of the label distribution just before the label is sampled — (conversation) calibration bounds therefore follow from standard Martingale concentration arguments.
2. Next, we observe that if two Bayesians are interacting with one another in the sequential setting, and the instance is drawn i.i.d. at each day, then the conversation that they have at each day t is statistically independent of all previous days. We know (from part 1) that if we allow them to interact across sufficiently many days, the transcript of their conversations will be arbitrarily well conversation calibrated, and hence in the canonical setting, they will agree on a $1-\delta$ fraction of days after $k = 1/\epsilon^2\delta$ many rounds. Similar guarantees with different bounds hold in each of our other settings. However, because the conversations at each round are identically and independently distributed, the transcript distribution is permutation invariant — and hence the two Bayesians will agree on the *first* day after at most $k = 1/\epsilon^2\delta$ many rounds, with probability $1-\delta$ over the selection of a day from the transcript, which is equivalent to a $1-\delta$ probability guarantee over the draw of the instance from the underlying prior.

Hence we conclude that our theorems extend to the 1-shot Bayesian setting and generalize and extend past work on Bayesian agreement. In particular we give quantitative convergence bounds in the style of Aaronson [2005] that are independent of the complexity of the instance, but are able to recover theorems not just in the cannonical setting, but in the d-dimensional and action feedback settings as well.

7.1 Bayesians are Conversation Calibrated

In this section we begin by showing that if the instance at each day is drawn from a prior distribution $\mathcal{D}$, and one of the Agents is a Bayesian who correctly computes predictions as posterior expectations given the prior $\mathcal{D}$ and all observed evidence, then when interacting with any other agent, they are guaranteed to maintain conversations that satisfy any of our calibration conditions. We start by defining how a Bayesian learner interacts in a conversation.

Definition 7.1 (Bayesian Learner). *Fix a prior $\mathcal{D} \in \Delta(\mathcal{X}_h \times \mathcal{X}_m \times \mathcal{Y})$ specifying a joint distribution over features observable to both the human and the model and labels. We say that a human (respectively, model) is a Bayesian Learner with prior $\mathcal{D}$ if given a known algorithm for the model, for all $t, k > 0$, given observable features x^t, message transcript $\mu^{1:t-1}$, prediction transcript $\pi_h^{1:t-1}$ of human predictions (respectively, $\pi_m^{1:t-1}$ of model predictions) through day $t-1$, and conversation $C^t_{1:k-1}$, they make a prediction as*

$$\hat{y}_h^{t,k} = \mathbb{E}_{\mathcal{D}}[Y|x^t, \mu^{1:t-1}, \pi_h^{1:t-1}, C^t_{1:k-1}] \quad \textit{(respectively, } \hat{y}_m^{t,k} = \mathbb{E}_{\mathcal{D}}[Y|x^t, \mu^{1:t-1}, \pi_m^{1:t-1}, C^t_{1:k-1}]).$$

[ht]

```
Input $(\mathcal{D}, \Omega_h, \Omega_m, \mathcal{Y}, \textsc{Agree}_\epsilon)$
for each day $t = 1, \dots$ do
  Receive $x^t = (x^t_h, x^t_m, y^t) \sim \mathcal{D}$. The model sees $x^t_m$ and the human sees $x^t_h$.
  for each round $k = 1, 2, \dots, L$ do
    if $k$ is odd then
      The Model predicts $\hat{y}_m^{t,k} \in \mathcal{Y}$, and sends the Human $p_m^{t,k} \in \Omega_m$
      if $\textsc{Agree}_\varepsilon(p_h^{t,k-1}, \hat{y}_h^{t,k-1}, p_m^{t,k}, \hat{y}_m^{t,k})$ then
        Return $p_m^{t,k}$ and break out of loop
    if $k$ is even then
      The Human predicts $\hat{y}_h^{t,k}$, and sends the model $p_h^{t,k} \in \Omega_h$
      if $\textsc{Agree}_\varepsilon(p_h^{t,k}, \hat{y}_h^{t,k}, p_m^{t,k-1}, \hat{y}_m^{t,k-1})$ then
        Return $p_m^{t,k-1}$ and break out of loop
  The Human and Model observe $y^t \in \mathcal{Y}$
```

Protocol 7.1 is the same as our general agreement protocol (Protocol 3.1), except that the instance at each day t is drawn i.i.d. from a prior distribution $\mathcal{D}$, rather than being chosen by an adversary. We will prove the following theorem, which states that if the human is a Bayesian learner, then they will satisfy strong conversation calibration constraints of various forms.

Theorem 7.2. *Consider an interaction over T rounds under Protocol 7.1. If the human (respectively, model) is a Bayesian Learner (Definition 7.1), then for any model (respectively, human) algorithm, for any $n > 0$, with probability $1 - \delta$, they are*

- $\left(O(T^{\frac{3}{4}}(\log\frac{dn}{\delta})^{\frac{1}{4}}), \frac{1}{n}\right)$*-conversation calibrated, and*
- $\left(2\sqrt{2T\log\frac{d|\mathcal{A}|^2}{\delta}}\right)$*-DC-conversation-calibrated.*

First we formalize a simple observation in the following lemma. It states that if we resample the label every day after the j^{th} round of conversation *from the posterior distribution on the label conditional on the transcript of interaction so far*, that this does not change the distribution of transcripts. An upshot of this lemma is that all of our subsequent analysis can proceed under this resampling thought experiment.

Lemma 7.3. *Let $\mathcal{D}$ be a probability distribution over space $\mathcal{X}_m \times \mathcal{X}_h \times \mathcal{Y}$ and fix a day $t \in [T]$. Fix a transcript through day $t-1$: $\pi^{1:t-1}$.*

- *Consider an interaction at day t under Protocol 7.1. Let π^t be the transcript of day t from this interaction.*

- *Fix an arbitrary round j. Consider an interaction when (x_m, x_h, y^t) is sampled from $\mathcal{D}$ at the beginning of day t and then the human and model correspond according to Protocol 7.1 until round j. Then, in round j, the outcome is resampled from the posterior distribution conditional on the information observed by the human so far: $y' \sim \mathcal{D}_\mathcal{Y}|x_h^t, \mu^{1:t-1}, \pi_h^{1:t-1}, C_{1:j-1}^t, p_m^{t,j}$. Let $\bar{\pi}_j^t$ be the transcript of day t from this interaction, with y^t replaced with y'.*

For all rounds k,

$$\mathop{\mathbb{P}}_{\mathcal{D}}[\pi^{t,1:k}] = \mathop{\mathbb{P}}_{\mathcal{D}}[\bar{\pi}_j^{t,1:k}].$$

The proof can be found in Appendix D

Lemma 7.3 tells us that we can proceed in our analysis by imagining that at any round j on which the Bayesian learner sends a message, they send a message that is consistent with the *true* label expectation at that round, as we can imagine that the label is resampled according to its posterior expectation. This means that the Bayesian's forecasts are unbiased, and so by Azuma's inequality, the average of the Bayesian's forecasts should equal the average of the realized label up to small error terms on any sequence that is sufficiently long. Thus the rest of the analysis consists of identifying sufficiently long sequences on which bounding the bias of the Bayesian's predictions in this way is sufficient to bound each notion of calibration error. This is enough to straightforwardly give us a bound on the Bayesian's decision conversation calibration error, since DC-conversation-calibration error is simply the maximum bias in any coordinate of the learner's predictions conditional on the best response action defined by the Bayesian's prediction and the action communicated by the other agent; thus there are only $|\mathcal{A}|^2$ many sequences on which we need to bound the bias, and the result will follow from Azuma's inequality and a union bound. However, conversation calibration is defined in terms of *distance to calibration*, which is more subtle. Distance to calibration is upper bounded by expected calibration error (ECE), however the empirical ECE of a Bayesian will in general *not* be bounded, as they might make a different prediction at every round, and hence there will be no sequences of fixed predictions of length > 1, and hence we have no ability to invoke concentration. Instead, we will bound the Bayesian's *bucketed expected calibration error*, defined next, and use this to upper bound distance to calibration.

Definition 7.4 (Bucketed Expected Calibration Error)**.** *Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, the expected calibration error with respect to bucketing coarseness n (Definition 3.21) is*

$$\mathrm{ECE}(p^{1:T}, y^{1:T}; n) = \sum_{i=1}^{n} \left| \sum_{t=1}^{T} \mathbb{1}[p^t \in B_n(i)](p^t - y^t) \right|.$$

Lemma 7.5. *Fix a sequence of of predictions $p^{1:t}$ and outcomes $y^{1:T}$. Then,* $\mathrm{CalDist}(p^{1:T}, y^{1:T}) \leq \mathrm{ECE}(p^{1:T}, y^{1:T}; n) + \frac{T}{n}$.

The proof is in Appendix D.

Bounding bucketed calibration error for a Bayesian can be done via Azuma's inequality: it now reduces to bounding the empirical bias of the predictions conditional on the bucket of the prediction, which for a bucketing parameter n consists of n subsequences, each of which we can apply Azuma's inequality to. The final bounds come from optimizing n, trading off the need to sum over the magnitude of the bias on each sequence defined by a bucketing (which is costlier for larger n) and the need to bound distance to calibration using Lemma 7.5 (which is costlier for smaller n). The details are in Appendix D.

7.2 An Online to One-Shot Reduction

In this section, we show that if an instance is drawn from a commonly known prior, and *both* agents are Bayesian, then all of our theorems that bound the conversation length K for a $1-\delta$ fraction of conversations over an arbitrarily long sequence of length T in fact hold for a *single* conversation, with probability $1-\delta$ over the draw of the instance from the prior distribution. The idea is straightforward: We can *imagine* an arbitrarily long sequence of conversations over many days. Because we showed that Bayesians satisfy our notions of conversation calibration with parameters growing sublinearly with T, our theorems apply with the error terms going to 0 as T grows large, and we can conclude that their conversations are short for a $1-\delta$ fraction of days. But we can also observe that because the instances are drawn i.i.d. from a fixed prior, and in such a setting Bayesians need not condition on any information from prior days, the conversation on each day is distributed identically. Hence it must be that *each* conversation (and in particular the first) is bounded with probability $1-\delta$ over the prior. We therefore conclude that our theorems hold for a single conversation between Bayesians.

[ht]

```
Input $(\Omega_h, \Omega_m, \mathcal{Y}, \textsc{Agree}_\epsilon, \mathcal{D} \in \Delta(\mathcal{X}_h \times \mathcal{X}_m \times \mathcal{Y})$, instance for which you want agreement: $(x_h^*, x_m^*, y^*) \sim \mathcal{D}$
Parameter agreement tolerance: $\varepsilon$, failure probability: $\delta$, number of samples: $T$
Let $(x_h^1, x_m^1, y^1) = (x_h^*, x_m^*, y^*)$
For $t \in \{2, \dots, T\}$ draw $(x_h^t, x_m^t, y^t) \sim \mathcal{D}$
for each day $t = 1, \dots, T$ do
  Model observes $x_m^t$ and Human observes $x_h^t$.
  for each round $k = 1, 2, \dots, L$ do
    if $k$ is odd then
      The Model predicts $\hat{y}_m^{t,k} \in \mathcal{Y}$, and sends the Human $p_m^{t,k} \in \Omega_m$
      if $\textsc{Agree}_\varepsilon(p_h^{t,k-1}, \hat{y}_h^{t,k-1}, p_m^{t,k}, \hat{y}_m^{t,k})$ then
        Return $p_m^{t,k}$ and break out of loop
    if $k$ is even then
      The Human predicts $\hat{y}_h^{t,k}$, and sends the model $p_h^{t,k} \in \Omega_h$
      if $\textsc{Agree}_\varepsilon(p_h^{t,k}, \hat{y}_h^{t,k}, p_m^{t,k-1}, \hat{y}_m^{t,k-1})$ then
        Return $p_m^{t,k-1}$ and break out of loop
  The Human and Model observe $y^t \in \mathcal{Y}$
```

We define a hypothetical conversation protocol (Protocol 7.2) that takes as input a prior distribution $\mathcal{D}$ and a single instance (x_h^*, x_m^*, y^*) drawn from the prior distribution that we want fast agreement on. The hypothetical protocol runs our agreement protocol for T rounds, using the supplied instance (x_h^*, x_m^*, y^*) on day 1, and using freshly sampled instances from the prior at all subsequent days. Note that we will never run Protocol 7.2 — in particular, in reality, we do not want to have to know the label y^* before the Bayesians converse — but it will be a useful thought experiment.

A fixed Human algorithm, denoted H, a fixed Model algorithm, denoted M, a prior $\mathcal{D}$, and hypothetical Protocol 7.2 together define a distribution over transcripts. Of particular interest to us will be the distribution over conversation lengths at each round. Let $\ell_t(H, M, \mathcal{D})$ represent the conversation length at round t of the transcript induced by H, M, and $\mathcal{D}$ in Protocol 7.2. We first observe that the conversation lengths are identically distributed at each day of Protocol 7.2, since the instances each day are i.i.d.:

Lemma 7.6. *If H and M are Bayesian learners, then $\mathbb{P}(\ell_{t_1}(H, M, \mathcal{D}) \geq k) = \mathbb{P}(\ell_{t_2}(H, M, \mathcal{D}) \geq k)$, $\forall t_1, t_2 \in [1, \dots, T], k \in \mathbb{N}$.*

Proof. Because the instance at each day t is drawn i.i.d., the predictions of a Bayesian Learner in round k are a function only of the prior $\mathcal{D}$, the feature vector they observe (x_h^t or x_m^t) and the conversation up to that round $C_{1:k-1}^t$. Therefore, given two Bayesian learners, $\ell_t(H, M, \mathcal{D})$ is a function only of (x_m^t, x_h^t, y^t). But (x_m^t, x_h^t, y^t) are i.i.d. for all t. Therefore, $\mathbb{P}(\ell_{t_1}(H, M, \mathcal{D}) \geq k) = \mathbb{P}(\ell_{t_2}(H, M, \mathcal{D}) \geq k)$, $\forall t_1, t_2 \in [1, \dots, T], k \in \mathbb{N}$. □

Next, we show that in the limit as the number of rounds T in the hypothetical Protocol 7.2 tends to infinity, we can give a high probability bound on the length K of the first conversation in Protocol 7.2 — i..e the conversation pertaining to the relevant instance (x_h^*, x_m^*, y^*). This follows because 1) Bayesians become increasingly conversation calibrated as T grows large, and so we can apply our theorems establishing that a $1-\delta$ *fraction* of conversations in Protocol 7.2 are short, and because 2) all conversation lengths are identically distributed, so if most conversations are short, it must also be that the *first* conversation is short with high probability.

Theorem 7.7. *Fix any $\epsilon, \delta \in [0, 1]$ and any instance $(x_h^*, x_m^*, y^*) \sim \mathcal{D}$. If the Human and the Model are both Bayesian learners, then under Protocol 7.2, in the limit as $T \to \infty$ they will reach*

$\epsilon-$agreement with probability $1-\delta$ on day 1 (i.e. the day corresponding to the instance (x_h^*, x_m^*, y^*)) within

- *$K \leq \frac{3d}{\epsilon^2\delta}$ rounds in the full feedback setting.*
- *$K \leq \frac{3}{2\varepsilon\delta} + 1$ rounds in the action feedback setting.*

Proof of Theorem 7.7. **The Full Feedback Setting:** By Theorem 7.2, if we run the protocol for T rounds, then with probability $1 - 2\delta/3$ both the Human and the Model are $\left(2\left(2T\log(\frac{3dT^{3/7}}{\delta})\right)^{\frac{1}{4}}, T^{-\frac{3}{7}}\right)$-conversation-calibrated. Assume for now that these calibration bounds hold.

Note that Protocol 7.2 is simply a special case of Protocol 3.1, in which (x_h^t, x_m^t, y^t) are drawn from a fixed distribution. Therefore, the guarantees from Theorem 5.2 hold, and we have that, any $\epsilon, \delta \in [0,1]$, on a $1 - \delta/3$ fraction of days, they reach ε-agreement after at most K rounds of conversation for $K \leq \frac{3d}{\epsilon^2\delta - \beta(T)}$ and where $\beta(T) = 3d\left(2T^{-\frac{3}{7}} + \frac{4(2T\cdot T^{-\frac{3}{7}}\log(\frac{3dT^{\frac{3}{7}}}{\delta}))^{\frac{1}{4}}}{T\cdot T^{-\frac{3}{7}}}\right)$. But $\lim_{T\to\infty}\beta(T) = 0$, and so we have that for every $\eta > 0$, $K < \frac{3d}{\epsilon^2\delta - \eta}$ Therefore we must have $K \leq \frac{3d}{\epsilon^2\delta}$.

Now, note that by Lemma 7.6, the distribution over conversation lengths at each day is identical. Therefore, we have that

$$\mathop{\mathbb{P}}_{t\sim Unif(1:T)}\left[\ell_t \geq \frac{3d}{\epsilon^2\delta}\right] \leq \frac{\delta}{3} \implies \mathbb{P}\left[\ell_1 \geq \frac{3d}{\epsilon^2\delta}\right] \leq \frac{\delta}{3}$$

Summing up all three failure probabilities, we have that

$$\mathbb{P}\left[\ell_1 \geq \frac{3d}{\epsilon^2\delta}\right] \leq \delta$$

The Action Feedback Setting By Theorem 7.2, if we run the protocol for T rounds, the human and model are both $(2\sqrt{2T\log\frac{3d|A|^2}{\delta}})$-decision-conversation calibrated with probability $1 - \frac{2\delta}{3}$. Assume for now these two calibration bounds hold. We instantiate Theorem 6.2: the human and model will reach $\varepsilon-$agreement on a $1 - \delta/3$ fraction of days, after at most

$$K \leq \frac{1}{2\epsilon\frac{\delta}{3} - \gamma(T)} + 1$$

rounds of conversation, where $\gamma(T) = \frac{4Ld|A|^2\cdot\sqrt{2T\log\frac{3d|A|^2}{\delta}}) + 4Ld|A|^2\cdot\sqrt{2T\log\frac{3d|A|^2}{\delta}})}{T}$. Here $\lim_{T\to\infty}\gamma(T) = 0$. So once again we have that for every $\eta > 0$, $K < \frac{1}{2\epsilon\frac{\delta}{3} - \eta} + 1$. Hence it must be that: $K \leq \frac{3}{2\epsilon\delta} + 1$.

Now, note that by Lemma 7.6, the distribution over conversation lengths at each day is identical. Therefore, we have that

$$\mathop{\mathbb{P}}_{t\sim Unif(1:T)}\left[\ell_t \geq \frac{3}{2\epsilon\delta} + 1\right] \leq \frac{\delta}{3} \implies \mathbb{P}\left[\ell_1 \geq \frac{3}{2\epsilon\delta} + 1\right] \leq \frac{\delta}{3}$$

Summing up over all three failure probabilities yields

$$\mathbb{P}\left[\ell_1 \geq \frac{3}{2\epsilon\delta} + 1\right] \leq \delta$$

□

Finally we note that since we have proven that agreement happens quickly with high probability over the draw of the instance from the prior on *the first round* of Protocol 7.2 in the limit as T grows large, but the interaction at round 1 is independent of T, there is no need to run the protocol for more than a single round — we have proven agreement theorems in the "one-shot" setting of prior work Aumann [1976], Geanakoplos and Polemarchakis [1982], Aaronson [2005], Frongillo et al. [2023].

Corollary 7.8. *Fix any $\epsilon, \delta \in [0, 1]$ and any instance $(x_h^*, x_m^*, y^*) \sim \mathcal{D}$. If the Human and the Model are both Bayesian learners, then under Protocol 7.2 with $T = 1$, they will reach $\epsilon-$agreement on the instance (x_h^*, x_m^*, y^*), with probability $1 - \delta$ after at most*

- $K \leq \frac{3d}{\epsilon^2 \delta}$ *rounds in the full feedback setting.*
- $K \leq \frac{3}{2\varepsilon\delta} + 1$ *rounds in the action feedback setting.*

8 Discussion and Conclusion

Bayesian rationality is an attractive, canonical model of optimal learning that has been adopted in many economic models, including not just agreement (as we study in this paper), but also Bayesian Persuasion Kamenica and Gentzkow [2011], reputation systems Mailath and Samuelson [2006], and social herding Banerjee [1992]. While attractive, Bayesian reasoning is not computationally or statistically tractable, and so models that assume perfect Bayesian agents are either limited to speaking of extremely simple prior distributions or require making implausible assumptions on the knowledge and computational power of the agents. Motivated in part by these concerns, there is also a large literature that studies learning under simple behavioral assumptions (dating back to Simon [1955], Tversky and Kahneman [1992]) — but these models are generally incompatible with Bayesian reasoning, and hence are inherently less canonical — they require making choices about how to model agent behavior that have no firm theoretical grounding.

Our work suggests a third approach: We make computationally and statistically tractable calibration assumptions that are strict relaxations of Bayesian rationality, and hence are satisfied by perfect learners, but do not require implausible assumptions. In the case of agreement theorems, we have shown that these tractable calibration conditions were *all that was needed* from Bayesian rationality, in that we are able to prove (and generalize) agreement theorems that recover the same quantitative bounds that were known under full Bayesian rationality under our weaker assumptions. Is this a more general phenomenon? Perhaps in many other settings in which Bayesian rationality was previously thought to be a necessary modeling assumption, the same results can be obtained under significantly weaker calibration-based assumptions that can be guaranteed by efficient online calibration algorithms of various flavors.

Acknowledgments

We thank Joe Halpern and Rakesh Vohra for useful comments. We also thank Kavya Ravichandran for helpful discussions on human AI collaboration. This work was supported in part by the Simons Collaboration on the Theory of Algorithmic Fairness, NSF grants FAI-2147212 and CCF-2217062, the Hans Sigrist Prize, and an OpenAI SuperAlignment Fast Grant.

A Additional Material from Section 4

Lemma A.1. *If $m = \frac{1}{T}\sum_{t=1}^{T} y^t$, then for any constant x,*

$$\mathrm{SQE}(x, y^{1:T}) - \mathrm{SQE}(m, y^{1:T}) = \sum_{t=1}^{T}(x-m)^2 \tag{7}$$

Proof.

$$\begin{aligned}
\mathrm{SQE}(x, y^{1:T}) - \mathrm{SQE}(m, y^{1:T}) &= \sum_{t=1}^{T}(x-y_t)^2 - \sum_{t=1}^{T}(m-y_t)^2 \\
&= \sum_{t=1}^{T}(m-y_t+x-m)^2 - \sum_{t=1}^{T}(m-y_t)^2 \\
&= \sum_{t=1}^{T}(m-y_t)^2 + \sum_{t=1}^{T}(x-m)^2 + \sum_{t=1}^{T}2(m-y_t)(x-m) - \sum_{t=1}^{T}(m-y_t)^2 \\
&= \sum_{t=1}^{T}(x-m)^2 + \sum_{t=1}^{T}2(m-y_t)(x-m) \\
&= \sum_{t=1}^{T}(x-m)^2 + 2(x-m)\sum_{t=1}^{T}(m-y_t) \\
&= \sum_{t=1}^{T}(x-m)^2
\end{aligned}$$

□

Lemma A.2. *Let $T_k^{i,p_h} = \{t : p_h^{t,k} = p_h \text{ and } p_m^{t,k-1} \in B_i(\frac{1}{g(T)})\}$ be the subsequence of days such that the human predicts p_h in round k and the model predicts in bucket $B_i(\frac{1}{g(T)})$ in round $k-1$. Let $m_k^{i,p_h} = \frac{\sum_{t \in T_k^{i,p_h}} y^t}{|T_k^{i,p_h}|}$ be the true mean on this subsequence. If the human is $(\cdot, g_h(T))$-conversation calibrated, then*

$$\sum_{t \in T_k^{i,p_h}} (p_m^{t,k-1} - y^t)^2 - \sum_{t \in T_k^{i,p_h}} (i \cdot g_h(T) - y^t)^2 \geq -g_h(T) \cdot |T_k^{i,p_h}| \tag{8}$$

Proof. Note that for any t such that $\ell_t \geq k$, $(i-1) \cdot g_h(T) \leq p_m^{t,k-1} \leq i \cdot g_h(T)$, by the human's bucketing condition. Therefore, we also have that $(p_m^{t,k-1})^2 \geq ((i-1)g_h(T))^2$.

$$
\begin{aligned}
&\sum_{t \in T_k^{i,p_h}} (p_m^{t,k-1} - y^t)^2 - \sum_{t \in T_k^{i,p_h}} (i \cdot g_h(T) - y^t)^2 \\
&= \sum_{t \in T_k^{i,p_h}} \left((p_m^{t,k-1})^2 - 2p_m^{t,k-1} y^t + (y^t)^2\right) - \sum_{t \in T_k^{i,p_h}} \left((i \cdot g_h(T))^2 - 2(i \cdot g_h(T)) y^t + (y^t)^2\right) \\
&\hfill \text{(Expanding)} \\
&= \sum_{t \in T_k^{i,p_h}} (p_m^{t,k-1})^2 - (i \cdot g_h(T))^2 - 2p_m^{t,k-1} y^t + 2(i \cdot g_h(T)) y^t \qquad \text{(Cancelling out the } (y^t)^2\text{)} \\
&\geq \sum_{t \in T_k^{i,p_h}} ((i-1) g_h(T))^2 - (i \cdot g_h(T))^2 - 2p_m^{t,k-1} y^t + 2(i \cdot g_h(T)) y^t \\
&\hfill \text{(As } (p_m^{t,k-1})^2 \geq ((i-1) g_h(T))^2\text{)} \\
&\geq \sum_{t \in T_k^{i,p_h}} ((i-1) g_h(T))^2 - (i \cdot g_h(T))^2 - 2(i \cdot g_h(T))) y^t + 2(i \cdot g_h(T)) y^t \\
&\hfill \text{(As } p_m^{t,k-1} \leq i \cdot g_h(T)\text{)} \\
&= \sum_{t \in T_k^{i,p_h}} ((i-1) g_h(T))^2 - (i \cdot g_h(T))^2 \\
&= \sum_{t \in T_k^{i,p_h}} \left((i-1)^2 - i^2\right) g_h(T)^2 \\
&= \sum_{t \in T_k^{i,p_h}} (1 - 2i) g_h(T)^2 \\
&\geq \sum_{t \in T_k^{i,p_h}} (1 - \frac{2}{g_h(T)}) g_h(T)^2 \qquad \text{(As } i \leq \tfrac{1}{g_h(T)}\text{)} \\
&= \sum_{t \in T_k^{i,p_h}} (g_h(T)^2 - g_h(T)) \\
&\geq -|T_k^{i,p_h}| \cdot (g_h(T))
\end{aligned}
$$

□

Lemma A.3. *Consider any sequence of predictions and labels $p^{1:T}, y^{1:T}$ such that p is perfectly calibrated on y, and some other sequence of predictions $q^{1:T}$ such that $||p^{1:T} - q^{1:T}|| \leq \gamma$. Then,*

$$
\sum_{t=1}^{T} (q^t - y^t)^2 - \sum_{t=1}^{T} (p^t - y^t)^2 \leq 3\gamma
$$

Proof.

$$
\begin{aligned}
\sum_{t=1}^{T}(p^t - y^t)^2 - \sum_{t=1}^{T}\left(q^t - y^t\right)^2 &= \sum_{t=1}^{T}(p^t)^2 + (y^t)^2 - 2p^t y^t - \left(\sum_{t=1}^{T}(q^t)^2 + (y^t)^2 - 2q^t y^t\right) \\
&= \sum_{t=1}^{T}((p^t)^2 - (q^t)^2) + \sum_{t=1}^{T}(2q^t y^t - 2p^t y^t) \\
&\leq \sum_{t=1}^{T}((p^t)^2 - (q^t)^2) + 2\gamma \\
&\qquad\qquad\qquad\qquad \text{(as } \|p^{1:T} - q^{1:T}\| \leq \gamma \text{ and } y^t \in [0,1]) \\
&\leq \sum_{t=1}^{T}((p^t) - (q^t)) + 2\gamma \qquad\qquad \text{(as } p^t, q^t \in [0,1]) \\
&= 3\gamma.
\end{aligned}
$$

□

Protocol 4 AOST: Almost-One-Step-Ahead with unknown T

```
Input Sequence of outcomes $y^{1:T} \in \{0,1\}^T$, where $T$ is unknown a priori
Output Sequence of predictions $p^{1:T} \in \{0, \frac{1}{m}, ..., 1\}^T$ for some discretization parameter $m > 0$
$t \leftarrow 1$
$\bar{T} \leftarrow 1$
while $t \leq T$ do
  $t_0 \leftarrow t$
  $\bar{T} \leftarrow 2 \cdot \bar{T}$
  while $t \leq \bar{T}$ and $t \leq T$ do
    Given look-ahead predictions $\tilde{p}^{t_0:t-1}$, define the look-ahead bias conditional on a prediction
    $p$ as:
        $\alpha_{\tilde{p}^{1:t-1}}(p) := \sum_{s=t_0}^{t-1} \mathbb{I}[\tilde{p}^s = p](\tilde{p}^s - y^s)$
    Choose two adjacent points $p_i = \frac{i}{m}, p_{i+1} = \frac{i+1}{m}$ satisfying:
        $\alpha_{\tilde{p}^{t_0:t-1}}(p_i) \leq 0$ and $\alpha_{\tilde{p}^{t_0:t-1}}(p_{i+1}) \geq 0$
    Arbitrarily predict $p^t = p_i$ or $p^t = p_{i+1}$
    Upon observing the (adversarially chosen) outcome $y^t$, set look-ahead prediction
        $\tilde{p}^t = \underset{p \in \{p_i, p_{i+1}\}}{\arg\min} \ |p - y^t|$
    $t \leftarrow t + 1$
```

Proof of Theorem 4.7. Consider running Algorithm 4 with some sequence of outcomes $y^{1:T}$. Note that, by construction, when the algorithm terminates, $\frac{\bar{T}}{2} \leq T \leq \bar{T}$. We can upper bound the distance to calibration as

$$
\begin{aligned}
\mathrm{CalDist}(p^{1:T} y^{1:T}) &\leq \sum_{i=1}^{\log_2(\bar{T})-1} \mathrm{CalDist}(p^{2^{i-1}:2^i}, y^{2^{i-1}:2^i}) + \mathrm{CalDist}(p^{\frac{\bar{T}}{2}:T}, y^{\frac{\bar{T}}{2}:T}) \\
&\leq \sum_{i=1}^{\log_2(\bar{T})-1} (2\sqrt{2^{i-1}}+1) + (2\sqrt{T}+1) \\
&\qquad \text{(Algorithm runs a separate version of AOST for each } \bar{T}\text{, and by Theorem 4.6)} \\
&= \sum_{i=1}^{\log_2(\bar{T})-1} (2^{\frac{i+1}{2}}+1) + (2\sqrt{T}+1) \\
&= \log_2(\bar{T}) - 1 + \sum_{i=1}^{\log_2(\bar{T})-1} (2^{\frac{i+1}{2}}) + (2\sqrt{T}+1) \\
&= \log_2(\bar{T}) + 2\sqrt{T} + \sum_{i=1}^{\log_2(\bar{T})-1} ((\sqrt{2})^{i+1}) \\
&= \log_2(\bar{T}) + 2\sqrt{T} + (\sqrt{2})^2 \cdot \frac{\sqrt{2}^{\log_2(\bar{T})-1}-1}{\sqrt{2}-1} \leq \log_2(\bar{T}) + 2\sqrt{T} + 2 \cdot \frac{\sqrt{\bar{T}}-1}{\sqrt{2}-1} \\
&\qquad \text{(As this is a geometric series)} \\
&\leq \log_2(2T) + 2\sqrt{T} + 2 \cdot \frac{\sqrt{2T}-1}{\sqrt{2}-1} \\
&= O(\sqrt{T})
\end{aligned}
$$

□

B Additional Material from Section 5

Proof of Lemma 5.4. Let $T_k^{i,p_h}[j] = \{t : p_h^{t,k}[j] = p_h \textit{ and } p_m^{t,k-1}[j] \in B_i(\frac{1}{g(T)})\}$ be the subsequence of days such that the $p_h^{t,k}[j] = p_h$, and $p_m^{t,k-1}[j] \in B_i(\frac{1}{g(T)})$. Let $m_k^{i,p_h}[j] = \frac{\sum_{t \in T_k^{i,p_h}[j]} y^t}{|T_k^{i,p_h}[j]|}$ be the true mean on this subsequence. The difference in squared error of predictions in dimension j can be written as

$$
\begin{aligned}
&\sum_{t\in T_k^{i,p_h}[j]} (p_m^{t,k-1}[j]-y^t[j])^2 - \sum_{t\in T_k^{i,p_h}} (p_h^{t,k}[j]-y^t[j])^2 \\
&= \left[\sum_{t\in T_k^{i,p_h}} (p_m^{t,k-1}[j]-y^t[j])^2 - \sum_{t\in T_k^{i,p_h}[j]} (m_k^{i,p_h}[j]-y^t[j])^2\right] \\
&- \left[\sum_{t\in T_k^{i,p_h}[j]} (p_h^{t,k}[j]-y^t[j])^2 - \sum_{t\in T_k^{i,p_h}[j]} (m_k^{i,p_h}[j]-y^t[j])^2\right]
\end{aligned}
$$

(Adding and subtracting $\sum_{t\in T_k^{i,p_h}[j]}(m_k^{i,p_h}[j]-y^t[j])^2$)

$$
\begin{aligned}
&\geq \left[\sum_{t\in T_k^{i,p_h}[j]} (i\cdot g_h(T)-y^t[j])^2 - |T_k^{i,p_h}[j]|\cdot g_h(T) - \sum_{t\in T_k^{i,p_h}[j]} (m_k^{i,p_h}[j]-y^t[j])^2\right] \\
&\qquad - \left[\sum_{t\in T_k[j]^{i,p_h}} (p_h^{t,k}[j]-y^t[j])^2 - \sum_{t\in T_k[j]^{i,p_h}} (m_k[j]^{i,p_h}-y^t[j])^2\right]
\end{aligned}
$$

(By Lemma A.2)

$$
= \left[\sum_{t\in T_k[j]^{i,p_h}} (i\cdot g_h(T)-m_k[j]^{i,p_h})^2 - |T_k^{i,p_h}[j]|\cdot g_h(T)\right] - \left[\sum_{t\in T_k^{i,p_h}[j]} (p_h^{t,k}[j]-y^t[j])^2 - \sum_{t\in T_k^{i,p_h}[j]} (m_k^{i,p_h}[j]-y^t[j])^2\right]
$$

(By Lemma A.1)

$$
= \left[\sum_{t\in T_k[j]^{i,p_h}} (i\cdot g_h(T)-m_k[j]^{i,p_h})^2 - |T_k^{i,p_h}[j]|\cdot g_h(T)\right] - \left[\sum_{t\in T_k^{i,p_h}[j]} (p_h-y^t[j])^2 - \sum_{t\in T_k^{i,p_h}[j]} (m_k^{i,p_h}[j]-y^t[j])^2\right]
$$

(As by definition of $T_k^{i,p_h}[j]$, $p_h^{t,k}[j]=p_h$)

$$
\geq \left[\sum_{t\in T_k^{i,p_h}[j]} (i\cdot g_h(T)-m_k^{i,p_h}[j])^2 - |T_k^{i,p_h}[j]|\cdot g_h(T)\right] - \left[\sum_{t\in T_k^{i,p_h}[j]} (p_h-m_k^{i,p_h}[j])^2\right]
$$

(By Lemma A.1)

$$
\geq -|T_k^{i,p_h}[j]|\cdot g_h(T) + \sum_{t\in T_k^{i,p_h}[j]} (i\cdot g_h(T)-p_h)^2
$$

(As the human is $(0,g_h(T))$-conversation calibrated, $p_h = m_k^{i,p_h}[j]$)

Summing this up for all i, p_h:

$$
\begin{aligned}
&\sum_{\forall i,p_h}\left(-|T_k^{i,p_h}[j]|\cdot g_h(T) + \sum_{t\in T_k^{i,p_h}[j]} (i\cdot g_h(T)-p_h)^2\right) \\
&\geq -g_h(T)T + \sum_{\forall i,p_h}\sum_{t\in T_k^{i,p_h}[j]} (i\cdot g_h(T)-p_h)^2
\end{aligned}
$$

(As $g_h(T)$ is independent of i and p_h, and $\sum_{\forall i,p_h}\left|T_k^{i,p_h}[j]\right| \leq T$)

$$
\begin{aligned}
&\geq -g_h(T)T + \sum_{\forall i,p_h}\sum_{t\in T_k^{i,p_h}[j]} \mathbb{I}[|i\cdot g_h(T)-p_h^{t,k}[j]|\geq \epsilon - g_h(T)](i\cdot g_h(T)-p_h)^2 \\
&\geq -g_h(T)T + (\epsilon-g_h(T))^2\sum_{\forall i,p_h}\sum_{t\in T_k^{i,p_h}} \mathbb{I}[|i\cdot g_h(T)-p_h^{t,k}[j]|\geq \epsilon - g_h(T)]
\end{aligned}
$$

Note that, for all days in the subsequence $T_k^{i,p_h}[j]$, in round $k-1$ the model predicted in bucket $B_i(\frac{1}{g_h(T)}) = i \cdot g_h(T)$ in dimension j, and therefore in each of these days, by the definition of our bucketing, $p_m^{t,k-1}[j] \geq (i-1) \cdot g_h(T)$ and $p_m^{t,k-1}[j] \leq i \cdot g_h(T)$. So consider any round $t \in T_k^{i,p_h}[j]$. If $|p_h^{t,k}[j] - p_m^{t,k-1}[j]| \geq \epsilon$, then we have:

$$\begin{aligned}
|p_h^{t,k}[j] - p_m^{t,k-1}[j]| &\leq |p_h^{t,k}[j] - i \cdot g_h(T)| + |i \cdot g_h(T) - p_m^{t,k-1}[j]| \\
&= |p_h^{t,k}[j] - i \cdot g_h(T)| + i \cdot g_h(T) - p_m^{t,k-1}[j] \\
&\leq |p_h^{t,k}[j] - i \cdot g_h(T)| + i \cdot g_h(T) - (i-1) \cdot g_h(T) \\
&= |p_h^{t,k}[j] - i \cdot g_h(T)| + g_h(T), \\
\implies |p_h^{t,k}[j] - i \cdot g_h(T)| &\geq |p_h^{t,k}[j] - p_m^{t,k-1}[j]| - g_h(T) \geq \epsilon - g_h(T).
\end{aligned}$$

Thus, if $|p_h^{t,k}[j] - p_m^{t,k-1}[j]| \geq \epsilon$, then $|i \cdot g_h(T) - p_h^{t,k}[j]| \geq \epsilon - g_h(T)$, $\forall t \in T_k^{i,p_h}[j]$. Therefore the set of days for which the former condition holds is a subset of the latter condition, and we can write

$$\begin{aligned}
&- g_h(T)T + (\epsilon - g_h(T))^2 \sum_{\forall i, p_h} \mathbb{I}[|i \cdot g_h(T) - p_h| \geq \epsilon - g_h(T)] \cdot \left| T_k^{i,p_h}[j] \right| \\
&\qquad \geq -g_h(T)T + (\epsilon - g_h(T))^2 \sum_{\forall i, p_h} \sum_{t \in T_k^{i,p_h}[j]} \mathbb{I}[|p_h^{t,k}[j] - p_m^{t,k-1}[j]| \geq \epsilon]
\end{aligned}$$

Thus we have that

$$\begin{aligned}
&\sum_{\forall i, p_h} \left(\sum_{t \in T_k^{i,p_h}[j]} (p_m^{t,k-1}[j] - y^t[j])^2 - \sum_{t \in T_k^{i,p_h}} (p_h^{t,k}[j] - y^t[j])^2 \right) \\
&\qquad \geq -g_h(T)T + (\epsilon - g_h(T))^2 \sum_{\forall i, p_h} \sum_{t \in T_k^{i,p_h}[j]} \mathbb{I}[|p_h^{t,k}[j] - p_m^{t,k-1}[j]| \geq \epsilon]
\end{aligned}$$

Summing this up for all dimensions j:

$$\sum_{\forall j\in[d]}\left(\sum_{\forall i,p_h}\left(\sum_{t\in T_k^{i,p_h}[j]}(p_m^{t,k-1}[j]-y^t[j])^2-\sum_{t\in T_k^{i,p_h}}(p_h^{t,k}[j]-y^t[j])^2\right)\right)$$

$$\geq -dg_h(T)T+(\epsilon-g_h(T))^2\sum_{\forall j\in[d]}\left(\sum_{\forall i,p_h}\sum_{t\in T_k^{i,p_h}[j]}\mathbb{I}[|p_h^{t,k}[j]-p_m^{t,k-1}[j]|\geq\epsilon]\right)$$

$$=-dg_h(T)T+(\epsilon-g_h(T))^2\sum_{\forall i,p_h}\sum_{\forall j\in[d]}\sum_{t\in T_k^{i,p_h}[j]}\mathbb{I}[|p_h^{t,k}[j]-p_m^{t,k-1}[j]|\geq\epsilon]$$

$$\geq -dg_h(T)T+(\epsilon-g_h(T))^2\sum_{\forall i,p_h}\sum_{t\in T_k^{i,p_h}}\mathbb{I}[\exists j\in[d].s.t.|p_h^{t,k}[j]-p_m^{t,k-1}[j]|\geq\epsilon]$$

$$=-dg_h(T)T+(\epsilon-g_h(T))^2\sum_{t\in T_k}\mathbb{I}[\exists j\in[d].s.t.|p_h^{t,k}[j]-p_m^{t,k-1}[j]|\geq\epsilon]$$

$$\geq -dg_h(T)T+(\epsilon-g_h(T))^2|T^{\geq k+1}|$$

(As for every day that proceeds further than round k, there is ϵ disagreement in at least one coordinate)

As the human and the model are perfectly symmetrical, we also obtain the symmetrical result for the model. □

Proof of Theorem 5.2. By composing the two results in Theorem 5.5, until k such that $|T^{\geq k}|\leq\delta\cdot T$, we see that

$$\text{SQE}(p_h^{T^{\geq k},k-2},y^{T^{\geq k}})-\text{SQE}(p_h^{1:T,k},y^{T^{\geq k}})$$
$$\geq(\epsilon-g_h(T))^2|T^{\geq k+1}|+(\epsilon-g_m(T))^2|T^{\geq k}|-d\cdot(g_h(T)+g_m(T))T-3d\cdot(\frac{f_h(g_h(T)\cdot T)}{g_h(T)}+(\frac{f_m(g_m(T)\cdot T)}{g_m(T)})$$

Thus, consider any round r such that $|T^{\geq r}|\geq\delta T$. By applying this expression recursively, we can bound the squared error of the model at round r by

$$
\begin{aligned}
&\mathrm{SQErr}(p_h^{1:T,r}, y^{1:T,k}) \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - ((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2)\left(\sum_{k=1}^{r} |T^{\ge k}|\right) + d(g_m(T) + g_h(T))\left(\sum_{k=1}^{r} |T^{\ge k}|\right) \\
&\qquad + 3d\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)\left(\sum_{k=1}^{r} 1\right) \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - ((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2)\left(\sum_{k=1}^{r} |T^{\ge k}|\right) + d(g_m(T) + g_h(T))(r)T \\
&\qquad + 3d\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)(r) \qquad \text{(As } |T^{\ge k}| \le T\text{)} \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - ((\epsilon - g_m(T))^2 + (\epsilon - g_h(T))^2)(r)\delta T + 2d(g_m(T) + g_h(T))(r)T \\
&\qquad + 3d\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)(r) \\
&\qquad\qquad \text{(As for all } T^{\ge k} \text{ such that } k \le r,\ |T^{\ge k}| \ge \delta T\text{)} \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - (\epsilon^2)(r)\delta T + 3d(g_m(T) + g_h(T))(r)T \\
&\qquad + 3d\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)(r) \\
&\le \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - (r)\left((\epsilon^2)\delta T - 3d(g_m(T) + g_h(T))T - 3d\left(\frac{f_m(g_m(T)\cdot T)}{g_m(T)} + \frac{f_h(g_h(T)\cdot T)}{g_h(T)}\right)\right) \\
&= \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - (r)\left((\epsilon^2)\delta T - T\beta(T)\right)
\end{aligned}
$$

This completes the second part of the Theorem.

By definition, the squared error is non-negative. Therefore, we have that

$$
\begin{aligned}
&0 \le \mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k}) - r\left((\epsilon^2)\delta T - T\beta(T)\right) \\
&\implies r \le \frac{\mathrm{SQErr}(p_h^{T^{\ge r},1}, y^{1:T,k)}}{\epsilon^2\delta T - T\beta(T)} \\
&\implies r \le \frac{d\cdot T}{\epsilon^2\delta T - T\beta(T)} \\
&\implies r \le \frac{d}{\epsilon^2\delta - \beta(T)}
\end{aligned}
$$

This completes the first part of the Theorem. □

Proof of Theorem 5.5. Let $T_m(k,i,j) = \left\{t \in T^{\ge k} \mid p_m^{t,k-1}[j] \in B_i(1/g(T))\right\}$ be the subsequence of days where the j'th coordinate of the predictions of the model at round $k-1$ falls in bucket i and the conversation reaches round k. Note that by the definition of conversation calibration in d dimensions (Definition 3.23), we have that

$$
\mathrm{CalDist}(p_h^{T_m(k,i,j),k}[j], y^{T_m(k,i,j)}[j]) \le f(|T_m(k,i,j)|)
$$

Therefore, for predictions $p_h^{1:T,k}[j]$ from the human at round k in dimension j:

$$
\begin{aligned}
\mathrm{CalDist}(p_h^{1:T,k}[j], y^{1:T,k}[j]) &= \min_{q^{1:T} \in C(y^{1:T,k}[j])} \|p_h^{1:T,k}[j] - q_j^{1:T,k}\|_1 \\
&\qquad \text{(For 1-dimensional predictions } q_j\text{)} \\
&\leq \sum_{i=1}^{\frac{1}{g_h(T)}} \min_{q_j^{T_m(k,i,j)} \in C^{T_m(k,i,j)}(y^{1:T,k}[j])} \|p_h^{T_m(k,i,j),k}[j] - q_j^{T_m(k,i,j)}\|_1 \\
&\leq \sum_{i=1}^{\frac{1}{g_h(T)}} f_h(|T_m(k,i,j))|) \\
&\qquad \text{(By the calibration distance of the Human)} \\
&\leq \frac{f_h(g_h(T) \cdot |T^{\geq k}|)}{g_h(T)} \qquad \text{(By the assumption that } f_h \text{ is concave)} \\
&\leq \frac{f_h(g_h(T) \cdot T)}{g_h(T)}
\end{aligned}
$$

Let $q^{1:T,k,j}$ be a set of perfectly calibrated predictions that are $f_h(|T_m^{k,i,j}|)$-close to $p_h^{1:T,k}[j]$. Furthermore, let $q^{1:T,k}$ be the set of d-dimensional predictions such that $q^{1:T,k}[j] = q^{1:T,k,j}$. Then, we have,

$$
\begin{aligned}
\mathrm{SQErr}(p_h^{1:T,k}, y^{1:T}) &= \sum_{j \in [d]} \mathrm{SQErr}(p_h^{1:T,k}[j], y^{1:T}[j]) \\
&\leq \sum_{j \in [d]} \left(\mathrm{SQErr}(q^{1:T,k,j}, y^{1:T,k}[j]) + 3\frac{f_h(g_h(T) \cdot T)}{g_h(T)} \right) \qquad \text{(by Lemma A.3)} \\
&= \mathrm{SQErr}(q^{1:T,k}, y^{1:T,k}) + 3d \cdot \frac{f_h(g_h(T) \cdot T)}{g_h(T)} \\
&\leq \mathrm{SQErr}(p_m^{T^{\geq k},k-1}, y^{1:T,k}) - (\epsilon - g_h(T))^2 |T^{\geq k+1}| + d g_h(T) T + 3d\frac{f_h(g_h(T) \cdot T)}{g_h(T)} \\
&\qquad \text{(By Lemma 5.4)}
\end{aligned}
$$

As the Human and the Model are symmetric, we also obtain the symmetric result for the Model. □

Proof of Theorem 5.6. Algorithm 5 instantiates a copy of algorithm D for each round k, bucket i, and coordinate j pair corresponding to each of the sets $T_m(k,i,j)$. Therefore, we have that for each round k, bucket i, coordinate j,

$$
\mathrm{CalDist}(p_h^{T_m(k,i,j),k}[j], y^{T_m(k,i,j)}[j]) \leq f(|T_m(k,i,j)|)
$$

by assumption that D has worst-case distance to calibration $f_m(\cdot)$.

The fact that CONVERSE-dDIM$_1(x_m^t)[j] = M_0(x_m^t)[j]$, for all t follows by construction. □

C Additional Material from Section 6

Proof of Theorem 6.9. By Corollary 6.3, When $\gamma(T) \leq \varepsilon\delta$, on a $1 - \delta$ fraction of days, the number of rounds until agreement is at most

Protocol 5 CONVERSE-dDIM$(M_0, D, g_m(T))$: A reduction from an online decision-making algorithm to an algorithm with low conversation-calibration error in d dimensions

```
**Input** Baseline model algorithm $M_0$, D2C algorithm $D$, Discretization $g_m(T)$
We denote $D^j_{k,i}$ as an instantiation of $D$ which is given as input only the subsequence of days where $p_h^{t,k}[j] \in [(i-1)\cdot g_m(T), i \cdot g_m(T)]$, and denote $D^j_{k,i,t}$ be the prediction of $D^j_{k,i}$ at round $t$.
**for** $t = 1, \ldots, T$ **do**
  Receive $x_m^t$
  Send prediction $p_m^{t,1}[j] = M_0(x_m^t)[j]$ to human for each $j \in [d]$
  **for** $k = 3, 5, \ldots$ **do**
    **for** $j = 1, 2, \ldots, d$ **do**
      Initialize empty set $S$
      Observe human prediction $p_h^{t,k-1}[j]$
      **if** $|p_h^{t,k-1}[j] - p_m^{t,k-2}[j]| < \epsilon$ **then**
        Predict $p_h^{t,k-1}[j]$ and break out of loop
      Let $i$ be such that $p_h^{t,k-1}[j] \in [(i-1)\cdot g_m(T), i \cdot g_m(T)]$
      **if** $D^j_{k-1,i}$ uninitialized **then**
        Initialize $D^j_{k-1,i}$
      Send prediction $p_m^{t,k}[j] = D^j_{k-1,i,t}$ to human
      $S \leftarrow S \cup (k-1, i, j)$
      **if** $|p_h^{t,k-1}[j] - p_m^{t,k}[j]| < \epsilon$ **then**
        Predict $p_h^{t,k-1}[j]$ and break out of loop
  Observe $y^t$
  **for** $(k, i, j) \in S$ **do**
    Update $D^j_{k,i}$ with $(D^j_{k,i,t}, y^t)$
```

$K \leq \frac{1}{\varepsilon\delta} + 1$, where $\gamma(T) = \frac{2Ld|A|^2 \cdot f_h(\frac{T}{|A|^2}) + 2Ld|A|^2 \cdot f_m(\frac{T}{|A|^2})}{T}$. Instantiating this bound with the high-probability result for Theorem 6.8, we have that, with probability $1-\alpha$:

$$
\begin{aligned}
\gamma(T) &= \frac{4Ld|A|^2 \cdot O\left(\log(2d\,|\mathcal{E}|\,T) + \sqrt{T\ln(\frac{|\mathcal{E}|d}{\alpha})}\right)}{T} \\
&= \frac{4Ld|A|^2 \cdot O\left(\log(d\,|A|^2\,T) + \sqrt{T\ln(\frac{|A|^2 d}{\alpha})}\right)}{T} && \text{(By the definition of } \mathcal{E} \text{ in our setting)} \\
&= \frac{O\left(Ld|A|^2 \cdot \log(d\,|A|^2\,T) + Ld|A|^2 \cdot \sqrt{T\ln(\frac{|A|^2 d}{\alpha})}\right)}{T} \\
&= \frac{O\left(Ld|A|^2 \cdot \log(d\,|A|^2\,T)\right)}{T} + \frac{O\left(Ld|A|^2 \cdot \sqrt{\log(\frac{|A|^2 d}{\alpha})}\right)}{\sqrt{T}} \\
&\leq \frac{O\left(Ld|A|^2 \cdot \log(d\,|A|^2) + Ld|A|^2 \cdot \sqrt{\log(\frac{|A|^2 d}{\alpha})}\right)}{\sqrt{T}}
\end{aligned}
$$

Thus, to set $\gamma(T) \leq \epsilon\delta$ w.p. $\geq 1 - \alpha$, it is sufficient to set

$$
\begin{aligned}
& \frac{O\left(Ld|A|^2 \cdot \log(d\,|A|^2)\right)}{\sqrt{T}} + \frac{O\left(Ld|A|^2 \cdot \sqrt{\log(\frac{|A|^2 d}{\alpha})}\right)}{\sqrt{T}} \leq \epsilon\delta \\
\implies & \frac{O\left(Ld|A|^2 \cdot \log(d\,|A|^2) + Ld|A|^2 \cdot \sqrt{\log(\frac{|A|^2 d}{\alpha})}\right)}{\epsilon\delta} \leq \sqrt{T} \\
\implies & T \geq \frac{O\left(L^2 d^2 |A|^4 \cdot \log^2(d\,|A|^2) + L^2 d^2 |A|^4 \cdot \log(\frac{|A|^2 d}{\alpha})\right)}{\epsilon^2 \delta^2} \\
\implies & T \geq \frac{O\left(L^2 d^3 |A|^5 (1 + \log(\frac{1}{\alpha}))\right)}{\epsilon^2 \delta^2}
\end{aligned}
$$

□

D Additional Material from Section 7

D.1 Bayesians are Conversation Calibrated

Proof of Lemma 7.3. We want to show that for any round j when the resampling might occur, $\mathbb{P}_{\mathcal{D}}[\pi^{t,1:k}] = \mathbb{P}_{\mathcal{D}}[\bar{\pi}_j^{t,1:k}]$ for all k. For $k < j$, the claim follows immediately since there is no difference in the two sampling protocols. In round j, the claim that $\mathbb{P}_{\mathcal{D}}[\pi^{t,1:j}] = \mathbb{P}_{\mathcal{D}}[\bar{\pi}_j^{t,1:j}]$ follows from a generic statement about resampling from posterior distributions that we formalize in Lemma D.4: the joint distribution on any pair of random variables (A, B) is unchanged if we first sample a pair (A, B') and then sample B from its posterior distribution conditional on A. In this case, A is the distribution on the transcript $\pi^{t,1:j}$ excluding the label y^t and B is the label y^t. For rounds $k > j$ the claim holds since the distribution for the remaining interaction at round t is fixed once we fix $\pi^{t,1:k-1}$

□

Proof of Lemma 7.5. For any bucket $i \in [n]$, define the average outcome when the prediction falls in a bucket $B_n(i)$:

$$\bar{y}_i = \sum_{t=1}^{T} \frac{\mathbb{1}[p^t \in B_n(i)]}{\sum_{t'=1}^{T} \mathbb{1}[p^{t'} \in B_n(i)]} y^t$$

Similarly, let $\bar{p}_i$ define the average prediction in bucket i. Consider the sequence $q^{1:T}$ where $q^t = \bar{y}_i(p^t)$, where $p^t \in B_n(i)$. Observe that $q^{1:T}$ is perfectly calibrated.

$$
\begin{aligned}
\mathrm{CalDist}(p^{1:T}, y^{1:T}) &\leq \|p^{1:T} - q^{1:T}\|_1 \\
&= \sum_{t=1}^{T} |p^t - q^t| \\
&= \sum_{t=1}^{T} \sum_{i \in [n]} \mathbb{1}[p^t \in B_n(i)] |p^t - \bar{y}_i^t| \\
&\leq \sum_{t=1}^{T} \sum_{i \in [n]} \mathbb{1}[p^t \in B_n(i)] \left(|p^t - \bar{p}_i| + |\bar{p}_i - \bar{y}_i| \right) \quad \text{(by the triangle inequality)} \\
&= \sum_{i \in [n]} \sum_{t=1}^{T} \mathbb{1}[p^t \in B_n(i)] \left(|p^t - \bar{p}_i| \right) + \sum_{i \in [n]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t \in B_n(i)] \left(\bar{p}_i - \bar{y}_i \right) \right| \\
&\qquad \text{(by the fact that } \bar{p}_i \text{ and } \bar{y}_i \text{ are constant for each } i \in [n]) \\
&\leq \frac{T}{n} + \sum_{i \in [n]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t \in B_n(i)] \left(\bar{p}_i - \bar{y}_i \right) \right| \\
&\qquad \text{(by the fact that } |p - \bar{p}_i| \leq \tfrac{1}{n} \text{ for all } p \in B_n(i)) \\
&= \frac{T}{n} + \mathrm{ECE}(p^{1:T}, y^{1:T}; n).
\end{aligned}
$$

□

Proof of Theorem 7.2. Consider a modified interaction under Protocol 7.1 when, in each day in round j (if the conversation reaches round j), the outcome is resampled according to the information seen by the human so far: $y' \sim \mathcal{D}_{\mathcal{Y}} | x_h^t, \mu^{1:t-1}, \bar{\pi}_h^{1:t-1}, C_{1:j-1}^t, p_m^{t,j}$. Let $\hat{\pi}^j$ be the transcript from this interaction. First, we will show that $\mathbb{P}_{\mathcal{D}}[\pi] = \mathbb{P}_{\mathcal{D}}[\hat{\pi}^j]$, where π is the transcript under the unmodified Protocol 7.1.

Let $\hat{\pi}^{1:t,j}$ denote the transcript of this interaction up to day t. Note that this is distinct from $\bar{\pi}^{t,j}$, which denotes the transcript of an interaction only on day t where the resampling only occurs in round j. We will proceed via induction over days.

- **Base Case**: $\mathbb{P}_{\mathcal{D}}[\pi^{1:1}] = \mathbb{P}_{\mathcal{D}}[\hat{\pi}^{1:1,j}]$.

  Proof: On day $t = 1$, we have $\mathbb{P}_{\mathcal{D}}[\pi^1] = \mathbb{P}_{\mathcal{D}}[\bar{\pi}^{1,j}]$, by Lemma 7.3. Note that $\bar{\pi}^{1,j} = \bar{\pi}^{1:1,j} = \hat{\pi}^{1:1,j}$, and therefore $\mathbb{P}_{\mathcal{D}}[\pi^{1:1}] = \mathbb{P}[\hat{\pi}^{1:1,j}]$.

- **Inductive Step**: If $\mathbb{P}_{\mathcal{D}}[\pi^{1:t}] = \mathbb{P}_{\mathcal{D}}[\hat{\pi}^{1:t,j}]$, then $\mathbb{P}_{\mathcal{D}}[\pi^{1:t+1}] = \mathbb{P}_{\mathcal{D}}[\hat{\pi}^{1:t+1,j}]$.

  Proof: Observe that the state of the model algorithm in any round $t+1$ is a function only of the algorithm M and the transcript until that round: $\pi^{1:t}$ or $\bar{\pi}^{1:t}$. By the Inductive Hypothesis, $\mathbb{P}_{\mathcal{D}}[\pi^{1:t}] = \mathbb{P}_{\mathcal{D}}[\hat{\pi}^{1:t,j}]$ – and consequently, since the model algorithm M is the fixed between both interactions, therefore, $\mathbb{P}_{\mathcal{D}}[\pi^{t+1,j}] = \mathbb{P}_{\mathcal{D}}[\bar{\pi}^{t+1,j}]$. By Lemma 7.3, this is equal to $\mathbb{P}_{\mathcal{D}}[\pi^{t+1}]$. As $\mathbb{P}_{\mathcal{D}}[\hat{\pi}^{1:t,j}] = \mathbb{P}_{\mathcal{D}}[\pi^{1:t}]$ and $\mathbb{P}_{\mathcal{D}}[\bar{\pi}^{t+1,j}] = \mathbb{P}_{\mathcal{D}}[\pi^{t+1}]$, we have that $\mathbb{P}_{\mathcal{D}}[\pi^{1:t+1}] = \mathbb{P}_{\mathcal{D}}[\hat{\pi}^{1:t+1,j}]$.

Now, all that remains to show is that the human has low calibration error in transcript $\hat{\pi}(j)$. We will want to do so for each of our notions of conversation-calibration error.

Conversation-Calibration Error Fix some arbitrary round k. We will proceed by bounding the expected bucketed calibration error of the human conditioned on the model's previous message, and then applying Lemma 7.5 to show that this also bounds the human's conversation-calibration error.

Fix some bucketing coarseness m for the expected calibration error of the human and some bucket $v_h \in \mathcal{B}_m$. Fix some bucketing coarseness n and some bucket $v_m \in \mathcal{B}_n$.

We can then define a conditioning event $E : \Pi \to [0,1]$, defined as $E(\pi^{1:t}) = \mathbb{I}[p_h^{t,k} \in v_h, p_m^{t,k-1} \in v_m]$. Recall that the human is a Bayesian Learner. Therefore, their prediction is deterministic at the beginning of round k, since it is simply the posterior mean of the distribution conditioned on the model's predictions through round $k-1$.

Thus, we can instantiate Lemma D.3 with this event $E(\cdot)$: with probability $1-\delta$,

$$\left| \sum_{t=1}^{T} E(\pi^{1:t-1}) \cdot \left(y^t(\pi^{1:t-1})[j] - \mathbb{E}_{y \sim \mathcal{D}}[y[j] | \pi^{1:t-1}] \right) \right| \leq 2\sqrt{2T \ln \frac{1}{\delta}}.$$

Taking the union bound over all $j \in [d]$, $v_m \in \mathcal{B}_n$, we see that the magnitude of the bias of the human's predictions in coordinate j conditional on making a prediction in bucket v_h in round k is, with probability $1-\delta$, bounded by

$$2\sqrt{2T \ln \frac{dn}{\delta}}.$$

We can then sum across all buckets of the human's prediction $[m]$ to see that the expected calibration error of the human is bounded by

$$\text{ECE}(\hat{y}_h^{1:T,k} y^{1:T}; m) \leq \sum_{j \in [m]} 2\sqrt{2T \ln \frac{dn}{\delta}} = 2m\sqrt{2T \ln \frac{dn}{\delta}}.$$

Applying Lemma 7.5, we can bound the human's conversation-calibration error for fixed k, and all $i \in [n]$ and $j \in [d]$ as:

$$\text{CalDist}(\hat{y}_h^{T_m(k,i,j)}[j], y^{T_m(k,i,j)}[j]) \leq \frac{T}{m} + 2m\sqrt{2T \ln \frac{dn}{\delta}}.$$

Finally, setting the number of buckets $\mathcal{B}_m$ optimally as:

$$\frac{T}{m} = 2m\sqrt{2T \ln \frac{dn}{\delta}}$$

$$\downarrow$$

$$\frac{T}{2\sqrt{2T \ln \frac{dn}{\delta}}} = m^2$$

$$\frac{\sqrt{T}}{2(2T \ln \frac{dn}{\delta})^{1/4}} = m$$

$$\frac{T^{\frac{1}{4}}}{2(2 \ln \frac{dn}{\delta})^{\frac{1}{4}}} = m$$

We have a final bound of, for all $j \in [d], i \in \mathcal{B}_n$:

$$\text{CalDist}(\hat{y}_h^{T_m(k,i,j)}[j], y^{T_m(k,i,j)}[j]) \leq O(T^{\frac{3}{4}} (\ln \frac{dn}{\delta})^{\frac{1}{4}}).$$

We have shown this for an arbitrary round k. The claim that the human has low conversation-calibration error in the transcript $\hat{\pi}(k)$ holds for any round k when the resampling might occur, and so we have that with probability $1-\delta$, the human is $\left(O(T^{\frac{3}{4}} (\ln \frac{dn}{\delta})^{\frac{1}{4}}), \frac{1}{n} \right)$-conversation-calibrated.

DC-Calibration Error Fix some arbitrary round k. We proceed by bounding the magnitude of the bias of the predictions in each coordinate conditioned on the model's recommended action in the previous round and the best response to the human's prediction. Fix $a, a' \in \mathcal{A}$. We can then define a conditioning event $E : \Pi \to [0,1]$, where $E(\pi^{1:t}) = \mathbb{I}[p_m^{t,k-1} = a, p_h^{t,k} = a']$. Recall that the human is a Bayesian Learner. Therefore, their prediction is deterministic at the beginning of round

j, since it is simply the posterior mean of the distribution conditioned on the model's predictions through round $k-1$. Thus, we can instantiate Lemma D.3 with this event $E(\cdot)$ and see that with probability $1-\delta$,

$$|\sum_{t=1}^{T} E(\pi^{1:t-1}) \cdot \left(y^t(\pi^{1:t-1})[j] - \mathbb{E}_{y\sim\mathcal{D}}[y[j]|\pi^{1:t-1}]\right) | \leq 2\sqrt{2T\ln\frac{1}{\delta}}.$$

Taking the union bound over all $j \in [d], a, a' \in \mathcal{A}$, we see that the DC-calibration in round k is, with probability $1-\delta$, bounded by

$$2\sqrt{2T\ln\frac{d|A|^2}{\delta}}.$$

We have shown this for an arbitrary round k. The claim that the human has low DC-calibration error in the transcript $\hat{\pi}(k)$ holds for any round k when the resampling might occur, and so we have that, with probability $1-\delta$, the human is $(2\sqrt{2T\ln\frac{d|A|^2}{\delta}})$-DC-conversation-calibrated.

□

Theorem D.1 (Azuma's Inequality). *Let $\{X_0, X_1, \ldots\}$ be a martingale sequence such that $|X_{i+1} - X_i| < c$ for all i, then,*

$$\mathbb{P}[X_n - X_0 \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2c^2n}\right).$$

An immediate corollary of Theorem D.1 follows from appropriately setting parameters.

Corollary D.2. *Letting $X_0 = 0, \varepsilon = c\sqrt{2n\ln\frac{1}{\delta}}$, then we have for any $\delta \in (0,1)$, with probability $1-\delta$,*

$$X_n \leq c\sqrt{2n\ln\frac{1}{\delta}}.$$

Lemma D.3. *Let $E : \Pi \to [0,1]$ represent any conditioning event. Consider the random process $\{\mathcal{Z}^t\}$ adapted to the sequence of random variables π^t for $t \geq 1$ and let*

$$\mathcal{Z}^t := Z^{t-1} + E(\pi^{1:t-1}) \cdot \left(y^t(\pi^{1:t-1}) - \mathbb{E}_{y\sim\mathcal{D}}[y|\pi^{1:t-1}]\right)$$

Then,

$$\sum_{t=1}^{T} E(\pi^{1:t-1}) \cdot \left(y^t(\pi^{1:t-1}) - \mathbb{E}_{y\sim\mathcal{D}}[y|\pi^{1:t-1}]\right) \leq 2\sqrt{2T\ln\frac{1}{\delta}},$$

with probability $1-\delta$ over the randomness of $\mathcal{D}$ and $\pi^{1:t-1}$.

Proof. First, observe that the above sequence is a martingale as $\mathbb{E}_{\mathcal{D}}[E(\pi^{1:t-1}) \cdot (y^t(\pi^{1:t-1}) - \mathbb{E}_{y\sim\mathcal{D}}[y|\pi^{1:t-1}]] = E(\pi^{1:t-1}) \cdot \mathbb{E}_{\mathcal{D}}[(y^t(\pi^{1:t-1}) - \mathbb{E}_{y\sim\mathcal{D}}[y|\pi^{1:t-1}]] = 0$, since $E(\pi^{1:t-1})$ is a constant at the start day t as it does not depend on the outcome y^t. Thus, $\mathbb{E}_{\mathcal{D}}[Z^{t+1}] = Z^t$. Next, observe that since the outcomes $y \in [-1,1]$, we have the bounded difference condition: $|Z^t - Z^{t-1}| < 2$ for all t. We can then instantiate Azuma's Inequality with $n = T$ and $c = 2$ to get the claim. □

Lemma D.4 (Resampling). *Let $\mathcal{D}$ be a probability distribution over space $\mathcal{A} \times \mathcal{B}$. For all (a,b),*

$$\mathbb{P}_{(a,b)\sim\mathcal{D}, b'\sim\mathcal{D}|a}[(a,b')] = \mathbb{P}_{(a,b)\sim\mathcal{D}}[(a,b)].$$

Proof.

$$\begin{aligned}
\mathbb{P}_{(a,b)\sim\mathcal{D}, b'\sim\mathcal{D}|a}[(a,b')] &= \mathbb{P}_{a\sim\mathcal{A}}[a] \cdot \mathbb{P}_{b'\sim\mathcal{D}|a}[b'] \\
&= \mathbb{P}_{a\sim\mathcal{A}}[a] \cdot \mathbb{P}_{b\sim\mathcal{D}|a}[b] \\
&= \mathbb{P}_{(a,b)\sim\mathcal{D}}[(a,b)]
\end{aligned}$$

□

E Extension to Multiple Agents

In this section, we will extend our results to settings in which there are multiple agents interacting and aiming to reach agreement, rather than just two. For simplicity, we will restrict our attention here to the canonical setting (studied in Section 4), but our treatment here is meant to be exemplary: all of the settings we study can be efficiently extended to the n agent case in a similar manner. We will refer to the total number of agents as n. Since all agents in the canonical setting are symmetric (i.e. their calibration conditions and message spaces are the same), we will refer to them simply as agents in this section, rather than distinguishing between a specific number of humans and models. Informally, the results follows the same techniques as previously. We imagine a setting in which all n agents are *marginally* conversation calibrated with respect to the $n-1$ other agents. In fact, our results require only a weaker condition —- there should be some distinguished agent (agent 1) that satisfies $n-1$ marginal conversation calibration conditions with respect to his $n-1$ interlocutors, but the other agents only need to be conversation calibrated with respect to agent 1. Our algorithmic reduction will efficiently convert a model into one that can maintain all $n-1$ conversation calibration conditions simultaniously, so the algorithm can always serve the role of the distinguished agent, which allows us to make strictly weaker assumptions on the other parties. We note that once all agents $\epsilon/2$ agree with agent 1, they must also (by the triangle inequality) ϵ-agree with each other pairwise. Our analysis proceeds by showing that in any round in which an agent substantially disagrees with agent 1, the squared error of their predictions must improve relative to agent 1's predictions; similarly agent 1's predictions must improve relative to any other agent with which he disagrees substantially frequently.

We will first adapt our notation to handle n agents. We refer to the message space of an agent as Ω_a.

Definition E.1 (Agreement for n Agents)**.** *Given an agreement condition for two parties* $\textsc{Agree}$*, we define an agreement condition for* n *parties as the function:* $\textsc{n-Agree}_{\varepsilon,\textsc{Agree}} : (\Omega_a \times \mathcal{Y})^n \to \{0,1\}$ *defined as:*

$$\textsc{n-Agree}_{\varepsilon,\textsc{Agree}}(p_1, y_1, \ldots, p_n, y_n) = \begin{cases} 1, & \sum_{r \in \{2,\ldots,n\}} \textsc{Agree}_{\varepsilon/2}(p_1, p_1, p_r, y_r) = n-1 \\ 0, & \textit{otherwise.} \end{cases}$$

[ht]

```
Input $(\Omega_a, \mathcal{Y}, \textsc{Agree})$
for each day $t = 1, \ldots$ do
  Receive $x^t = (x^t_1, \ldots x^t_n)$. Agent $r$ sees $x^t_r$.
  for each round $k = 1, 2, \ldots, L$ do
    Set $i = k \mod n$
    Agent $i$ predicts $\hat{y}^{t,k}_i$ and sends all other agents $p^{t,k}_i \in \Omega_a$
    if $\textsc{n-Agree}_{\varepsilon,\textsc{Agree}}(p^{t,k}_i, \hat{y}^{t,k}_i, p^{t,k-1}_{i-1 \mod n}, \hat{y}^{t,k-1}_{i-1 \mod n}, \ldots, p^{t,k-(n-1)}_{i-(n-1) \mod n}, \hat{y}^{t,k-(n-1)}_{i-(n-1) \mod n}) = 1$ then
      Return $p^{t,k}_i$ and break out of loop
  Agents observe $y^t \in \mathcal{Y}$
```

We will need to slightly modify our conversation-calibration definitions to handle the general case of n agents. The idea is the same - an agent r is conversation-calibrated with respect to agent s if their predictions are calibrated conditional on the most recent message sent by agent s. The only difference is superficial - in the indexing of the subsequences of days of interest, which is made slightly more complicated by the introduction of multiple agents. In Protocol E, an agent r speaks in rounds k such that $k \equiv r \mod n$. For an arbitrary agent s, the most recent time they will have spoken prior to some round k (when an agent r is speaking) is: $k - ((r-s) \mod n)$.

Definition E.2 (Conversation-Calibrated Predictions with Many Agents)**.** *Fix an error function* $f : \{1, \ldots, T\} \to \mathbb{R}$ *and bucketing function* $g : \{1, \ldots, T\} \to (0, 1]$*. Given a prediction transcript* $\pi^{1:T}$ *resulting from an interaction in the canonical setting (Definition 3.4) with* n *agents, an agent* r *is* (f, g)*-conversation-calibrated with respect to agent* s *if for all rounds* $k \equiv r \mod n$ *and buckets* $i \in \{1, \ldots, 1/g(T)\}$*:*

$$\text{CalDist}(p^{T_s(k,i)}_r k, y^{T_s(k,i)}) \leq f(|T_s(k,i)|),$$

where $T_s(k,i) = \left\{ t \in T^{\geq k} \mid p^{t,k-((r-s) \mod n)}_s \in B_i(1/g(T)) \right\}$ *is the subsequence of days where the conversation reaches round* k *and the most recent prediction of agent* s *falls in bucket* i*.*

Theorem E.3. *If agent 1 is $(f(\cdot), g(\cdot))-$conversation-calibrated with respect to agents $2, \ldots, n$ and agents $2, \ldots, n$ are all $(f(\cdot), g(\cdot))-$conversation-calibrated with respect to agent 1, then: for any $\varepsilon, \delta \in [0, 1]$, on a $1 - \delta$ fraction of days, all agents reach $\varepsilon-$agreement after at most K rounds of conversation for*

$$K \le \frac{n}{\frac{\epsilon^2 \delta}{4} - \eta(T)},$$

where $\eta(T) = n\left(3g(T) + 6\frac{f(g(T))}{Tg(T)}\right)$.

Lemma E.4. *If agent 1 is marginally $(f(\cdot), g(\cdot))-$conversation-calibrated with respect to agents $2, \ldots, n$ and agents $2, \ldots, n$ are all $(f(\cdot), g(\cdot))-$conversation-calibrated with respect to agent 1, then for any round k such that $k = 1 \mod n$:*

$$\mathrm{SQE}(p_1^{1:T,k-n}, y^{T^{\ge k}}) \le \mathrm{SQE}(p_1^{1:T,k}, y^{T^{\ge k}}) - ((\frac{\varepsilon}{2}) - g(T))^2 \frac{|T^{\ge k+1}|}{n} + 2g(T)T + 6\frac{f(g(T))}{g(T)}$$

Proof of Lemma E.4. By the definition of Protocol E, if the conversation continued from round $k-n$ to k (for some $k \equiv 0 \mod n$), then every day, at least one other agent disagreed with agent 1's message in round $k-n$ by at least $\frac{\epsilon}{2}$. Then it must be the case that one of agents $2, \ldots, n$ disagreed with agent 1 in these rounds for at least $\frac{|T^{\ge k+1}|}{n}$ days. Then, the claim follows as a corollary to Lemma 4.4. □

Proof of Theorem E.3. Consider any round r such that $|T^{\ge r}| \ge \delta T$. We can create a telescoping sum by instantiating Lemma E.4 from round $k = 1$ to r:

$$\begin{aligned}
\mathrm{SQE}(p_1^{T^{\ge r},1}, y^{T^{\ge r}}) - \mathrm{SQE}(p_1^{T^{\ge r},r}, y^{T^{\ge r}}) &\ge \sum_{k=1}^{r} \left((\frac{\varepsilon}{2} - g(T))^2 |\frac{|T^{\ge r+1}|}{n}| - 2g(T)T - 6\frac{f(g(T))}{g(T)} \right) \\
&= r\left(\frac{1}{n}(\frac{\varepsilon}{2} - g(T))^2 \delta T - 2g(T)T - 6\frac{f(g(T))}{g(T)} \right)
\end{aligned}$$

Therefore:

$$\begin{gathered}
\mathrm{SQE}(p_1^{T^{\ge r},1}, y^{T^{\ge r}}) - r\left(\frac{1}{n}(\frac{\varepsilon}{2} - g(T))^2 \delta T - 2g(T)T - 6\frac{f(g(T))}{g(T)} \right) \ge 0 \\
\text{(as } \mathrm{SQE}(p_1^{T^{\ge r},r}, y^{T^{\ge r}}) \ge 0\text{)} \\
\downarrow \\
T - r\left(\frac{1}{n}(\frac{\varepsilon}{2} - g(T))^2 \delta T - 2g(T)T - 6\frac{f(g(T))}{g(T)} \right) \ge 0 \\
\text{(as } \mathrm{SQE}(p_1^{T^{\ge r},1}, y^{T^{\ge r}}) \le T\text{)} \\
\downarrow \\
\frac{T}{\frac{1}{n}(\frac{\varepsilon}{2} - g(T))^2 \delta T - 2g(T)T - 6\frac{f(g(T))}{g(T)}} \ge r
\end{gathered}$$

Hence:

$$\begin{aligned}
r &\le \frac{1}{\frac{\delta}{n}((\frac{\varepsilon}{2})^2 - 2\frac{\varepsilon}{2}g(T) + g(T)^2) - 2g(T) - 6\frac{f(g(T))}{Tg(T)}} \\
&\le \frac{1}{\frac{\delta}{n}(\frac{\varepsilon}{2})^2 - 3 \cdot g(T) - 6\frac{f(g(T))}{Tg(T)}} \\
&= \frac{n}{\frac{\delta\varepsilon^2}{4} - n \cdot (3 \cdot g(T) + 6\frac{f(g(T))}{Tg(T)})}.
\end{aligned}$$

□

We conclude with the algorithmic reduction. We will again use Theorem 6.6 and the framework introduced in Section 5. We will now have a different instantiation of the UnbiasedPrediction algorithm at each round k, and for the kth instantiation, the contexts each day t will be the conversation $C^{t,1:k-1}$ on that day so far.

We are interested in being unbiased in each round conditional on events which are defined marginally by each of the other agent's most recent bucketed prediction, and our own bucketed prediction. Thus, we define our event set accordingly. To do this, we will define a new bucketing set, $\hat{B}$, which has a different number of buckets $\frac{1}{g_1(T)}$. This will be the bucketing that we measure agent 1's bucketed ECE on, which we will then convert into a distance to calibration bound.

Definition E.5 (Multi-Conversation Events)**.** *For an agent s, a round k, and a pair of bucket indices i_1, i_2, let:*

$$E_{s,i_1,i_2,k}(x^t, \hat{y}^{t,k}, C^{t,1:k-1}) = \mathbb{1}\left[p_s^{t,k-((r-s) \mod n)} \in B_{i_1}\left(\frac{1}{g(T)}\right)\right] \mathbb{1}\left[p_1^{t,k} \in \hat{B}_{i_2}\left(\frac{1}{g_1(T)}\right)\right]$$

Let $\mathcal{E}_k := \{E_{s,i_1,i_2,k} \forall i_1, i_2, s\}$. Note that $|\mathcal{E}_k| = \frac{1}{g(T)} \cdot \frac{1}{g_1(T)} \cdot (n-1) \leq \frac{n}{g_1(T) \cdot g(T)}$.

We are now ready to define our reduction.

Protocol 6 CONVERSE-MANY(M_0, α)

```
Input Baseline model algorithm $M_0$, Discretization $g_m(T)$
for $t = 1, \ldots, T$ do
  Receive $x_m^t$
  Send prediction $p_m^{t,1} = M_0(x_m^t)$ to all other parties.
  for $k = 1, 1+n, 1+2n, \ldots$ do
    $L \leftarrow k$
    if $D_{k+1}$ uninitialized then
      Initialize $D_k = \text{UNBIASEDPREDICTION}(\mathcal{E}_k, \alpha)$
    Observe $n-1$ predictions $p_2^{t,k-((1-2) \mod n)}, \ldots, p_n^{t,k-((1-n) \mod n)}$
    if $\forall q, |p_q^{t,k-((1-q) \mod n)} - p_1^{t,k}| \leq \frac{\epsilon}{2}$ then
      Predict $p_1^{t,k}$ and break out of loop
    Send prediction $p = D_k(C^{t,1:k-1}, x_1^t)$
  Observe $y^t$
  for $k = 1, 1+n, 1+2n, \ldots L$ do
    Update $D_k$ with $y^t$
```

Theorem E.6. CONVERSE-MANY(M_0, α) *is* $(O(\ln(\frac{ndT}{g(T)g_1(T)}) + \sqrt{T\ln(\frac{nd}{\alpha g(T)g_1(T)})} + \frac{T}{g_1(T)}), g_1(T))$*-conversation-calibrated with respect to every other agent with probability $1-\alpha$, and for any sequence of labels $y^{1:T}$, its first-round prediction is the same as the prediction of the base model M_0 for all t:* CONVERSE-MANY$(M_0, \alpha)_1(x_m^t) = M_0(x_m^t)$*, for all t.*

Proof. By construction, in each round $1, n+1, \ldots, 2n+1$, CONVERSE-MANY(M_0, α) runs UNBIASEDPREDICTION$(\mathcal{E}_k, \alpha)$ with subsequences defined by $\mathcal{E}_k$ in order to obtain predictions. By Theorem 6.6, in each round and for each other agent s, the bias on subsequences defined by the other agent's bucketing and agent 1's bucketing is $O(\ln(\frac{ndT}{g(T)g_1(T)}) + \sqrt{T\ln(\frac{nd}{\alpha g(T)g_1(T)})})$. Note that by Lemma 7.5, agent 1's distance to calibration on this subsequence is therefore at most $O(\ln(\frac{ndT}{g(T)g_1(T)}) + \sqrt{T\ln(\frac{nd}{\alpha g(T)g_1(T)})} + Tg_1(T))$

Thus the algorithm is $(O(\ln(\frac{ndT}{g(T)g_1(T)}) + \sqrt{T\ln(\frac{nd}{\alpha g(T)g_1(T)})} + Tg_1(T)), g_1(T))$-conversation-calibrated with respect to every other agent.

The second result follows directly from the definition of CONVERSE-MANY(M_0, α). □

To be concrete about rates, we will assume for the purposes of the remaining Theorems that Agent 1 is employing Algorithm 6, and the remaining Agents are employing Algorithm 2. Thus, note by Corollary 4.2 that Agents $2, \ldots, n$ are $\sqrt{T}, T^{\frac{-1}{3}}$-conversation-calibrated, and therefore $g(T) = T^{-\frac{1}{3}}$.

Theorem E.7. *If $T \geq O\left(\frac{n^6}{\epsilon^{12}\delta^6} \cdot \ln^6(d)\right)$ and $g_1(T) = g^2(T)$, then with probability $\geq 1 - \alpha$, the number of rounds until agreement is at most*

$$K \leq \frac{2n}{\epsilon^2 \delta}$$

Proof. By Theorem E.3, when $\eta(T) \leq \varepsilon^2\delta$, on a $1 - \delta$ fraction of days, the number of rounds until agreement is at most

$K \leq \frac{n}{\frac{\varepsilon^2\delta}{4} - \eta(T)}$, where $\eta(T) = n\left(3g(T) + 6\frac{f(g(T))}{Tg(T)}\right)$. Instantiating this bound with the high-probability result for Theorem E.6, we have that, with probability $1 - \alpha$:

$$\begin{aligned}
\eta(T) &= n \cdot O\left(g(T) + \frac{\ln(\frac{ndT}{g(T)g_1(T)}) + \sqrt{T\ln(\frac{nd}{\alpha g(T)g_1(T)})} + Tg_1(T)}{Tg(T)}\right) \\
&= n \cdot O\left(T^{\frac{-1}{3}} + \frac{\ln(\frac{ndT}{T^{-1}}) + \sqrt{T\ln(\frac{nd}{\alpha T^{-1}})} + T^{\frac{1}{3}}}{T^{\frac{2}{3}}}\right) \\
&\qquad \text{(By the fact that } g_1(T) = g^2(T) \text{ and } g(T) - T^{\frac{-1}{3}}\text{)} \\
&= n \cdot O\left(T^{\frac{-1}{3}} + \frac{\ln(ndT^2) + \sqrt{T\ln(ndT)}}{T^{\frac{2}{3}}}\right) \\
&= n \cdot O\left(T^{\frac{-1}{3}} + \frac{\ln(nd) + \sqrt{T\ln(nd)}}{T^{\frac{2}{3}}}\right) \\
&= n \cdot O\left(\frac{\ln(nd)}{T^{\frac{2}{3}}} + \sqrt{\ln(nd)}T^{\frac{-1}{6}}\right) \\
&\leq n \cdot O\left(\ln(nd)T^{\frac{-1}{6}}\right)
\end{aligned}$$

Therefore, to set $K \leq \frac{2n}{\epsilon^2\delta}$, we can set

$$\begin{aligned}
&\frac{\epsilon^2\delta}{2} \geq n \cdot O\left(\ln(nd)T^{\frac{-1}{6}}\right) \\
&\implies T^{\frac{1}{6}} \geq O\left(\frac{2n}{\epsilon^2\delta} \cdot \ln(nd)\right) \\
&\implies T \geq O\left(\frac{n^6}{\epsilon^{12}\delta^6} \cdot \ln^6(nd)\right) \\
&\implies T \geq O\left(\frac{n^6}{\epsilon^{12}\delta^6} \cdot \ln^6(d)\right)
\end{aligned}$$

□