Output Agreement Mechanisms and Common Knowledge

Bo Waggoner
Harvard University
bwaggoner@seas.harvard.edu

Yiling Chen
Harvard University
yiling@seas.harvard.edu

Abstract

The recent advent of human computation – employing non-experts to solve problems – has inspired theoretical work in mechanism design for eliciting information when responses cannot be verified. We study a popular practical method, *output agreement*, from a theoretical perspective. In output agreement, two agents are given the same inputs and asked to produce some output; they are scored based on how closely their responses agree.

Although simple, output agreement raises new conceptual questions. Primary is the fundamental importance of *common knowledge*: We show that, rather than being truthful, output agreement mechanisms elicit common knowledge from participants. We show that common knowledge is essentially the best that can be hoped for in any mechanism without verification unless there are restrictions on the information structure. This involves generalizing truthfulness to include responding to a *query* rather than simply reporting a private signal, along with a notion of common-knowledge equilibria. A final important issue raised by output agreement is focal equilibria and player computation of equilibria. We show that, for eliciting the mean of a random variable, a natural player inference process converges to the common-knowledge equilibrium; but this convergence may not occur for other types of queries.

Portions of this work were presented at the *2013 Workshop on Social Computing and User-Generated Content*, at the 14th ACM Conference on Electronic Commerce.

Introduction

The emerging field of human computation has harnessed the intelligence of an unprecedentedly large population of people for the purpose of solving computational tasks. For example, in the now-classic ESP game (von Ahn and Dabbish 2004), which has collected semantic labels for over one hundred million images[1], the image labeling task is turned into a fun, online game: Two players are simultaneously shown an image and asked to independently type words related to the image; whenever a word is typed by both players, they score some points and move on to the next image.

The ESP game is an example of an *output agreement* mechanism, a term coined by von Ahn and Dabbish (2008) to describe a fundamental aspect of the game — rewarding agreement. While the ESP game has obtained an incredible amount of useful labels for images, it is interesting to ask what knowledge is elicited in such games with strategic players. Intuitively, a player will not always give the most descriptive label of an image in the ESP game if he thinks that that label may be too specialized to be known by the other player. For example, instead of "Woodcock", he may type "bird" for a picture of a Woodcock. Hence, we cannot expect to obtain all private knowledge of players in general. Then, exactly what knowledge can be reliably obtained?

This question motivates our effort in this paper. We formally define and analyze the broad class of output agreement mechanisms. In an output agreement mechanism, two players are presented with the same query and each gives a response, there is a metric measuring the distance (or degree of agreement) between the two responses, and the reward of the players monotonically decreases with the distance. For example, an output agreement mechanism can ask players to report some statistic of a random variable (e.g. the mean or median of customer ratings for a restaurant) and reward them according to the absolute difference of their reports. In this paper, we study what knowledge can be elicited at game-theoretic equilibria in output agreement mechanisms.

The output agreement mechanisms fall into the general setting that we refer to as *information elicitation without verification* (IEWV) because the designer would like to elicit useful information from the participants, but does not necessarily have the resources to verify the quality of responses. Many mechanisms have been developed for this setting, including the peer prediction method (Miller, Resnick, and Zeckhauser 2005) and Bayesian truth serum (Prelec 2004). However, the same model used for understanding prior mechanisms does not provide additional insights for output agreement beyond that it does not elicit all private knowledge. A theoretical analysis of output agreement requires novel approaches and insights that we believe are also relevant to understanding the broader IEWV setting as well.

In this paper, we first focus on the *solution concept*. Typically, mechanisms for IEWV ask agents to report their "signals", that is, the information they observe, and aim to truthfully elicit such signals under some assumptions on the structure of players' information or the mechanism's knowledge about it. But for output agreement, eliciting "signals" may

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] http://news.bbc.co.uk/2/hi/technology/7395751.stm

be unnecessary or infeasible. We model output agreement as asking agents a "query" and introduce a notion of player *specificity* to capture the amount or "coarseness" of knowledge that the player uses to answer the query. For example, "Woodcock" is a very specific response (it might be exactly the player's signal), while "small bird" is more coarse (though perhaps still useful), and "creature" is very coarse. Technically, the most refined knowledge that the player can use is his private signal (i.e. being truthful) while the coarsest knowledge is the prior information.

With this, we show that output agreement games elicit *common knowledge*: There is a strict equilibrium where players report the correct answer according to the common knowledge they possess; and this holds for *any* query we ask and any information structure agents have. We note that most prior mechanisms focus on only eliciting signals rather than arbitrary queries and often require assumptions on the information structure. Moreover, output agreement's solution of common knowledge cannot be much improved: No mechanism for IEWV can obtain answers that are based on strictly more refined knowledge (in particular, players' private information), without making restrictions on the structure of players' information. Another drawback of output agreement is the existence of "bad" equilibria where no information is revealed; we formalize this with *uninformative equilibria* and show that it is (virtually) impossible for a mechanism for IEWV to avoid this problem.

We second focus briefly on some of the implications of the common-knowledge solution concept on focal equilibria in output agreement. In prior mechanisms for IEWV, which focused on *truthful* equilibria, it might naturally be argued that such equilibria are focal: Agents are presented with a query and they respond truthfully. In output agreement, however, truthful responses are not always an equilibrium. If "Amanda" and "Ben" are playing an output agreement game, then Amanda may observe the query and think of a truthful response, but she must also reason about Ben's possible truthful responses and her own best response to these. But Ben should be following the same reasoning and should therefore best-respond to Amanda's best response; and so on.

Ideally, this *player inference process* would converge, by iterated computation of hypothetical best responses, to the common-knowledge equilibrium. We show that for reporting the mean of a random variable in $\mathbb{R}^n$, the inference process indeed converges to the common-knowledge equilibrium. But this is not the case for querying the median or mode of a random variable. Even if both players know that an outcome for a binary variable will happen almost certainly, hence this outcome is the median and mode, the inference process may converges to an equilibrium where both players always report the other outcome.

For brevity, in most cases cases our proofs will be omitted; they are available in the full version posted on the authors' webpages.

Related Work

Prior work in information elicitation without verification includes notably the peer prediction method (Miller, Resnick, and Zeckhauser 2005), its improved variants (Jurca and Faltings 2006; 2007a; 2009; 2007b) and Bayesian truth serum (Prelec 2004); these are most closely related to output agreement along with their extensions, peer prediction without a common prior (Witkowski and Parkes 2012b) and the robust Bayesian truth serum (Witkowski and Parkes 2012a; Radanovic and Faltings 2013). Other approaches focus on observations drawn i.i.d. from an unknown distribution in $\mathbb{R}$ (Lambert and Shoham 2008; Goel, Reeves, and Pennock 2009). Dasgupta and Ghosh (2013) design a mechanism to elicit binary evaluations when there are multiple simultaneous queries for each agent and agents can exert more effort to improve accuracy relative to an unknown ground truth.

The term "output agreement" was introduced by von Ahn and Dabbish (2008), with a primary example being the ESP Game (von Ahn and Dabbish 2004). Such games have been investigated experimentally (Weber, Robertson, and Vojnovic 2008; Huang and Fu 2012). But to our knowledge, there has been no theoretical analysis of the general output agreement setting. Witkowski et al. (2013) consider a very simple output agreement setting, but suppose there is an underlying (binary) truth to be discovered and that agents can invest additional effort to gain additional information about the truth. Jain and Parkes (2008) give a game-theoretic model and analysis of the ESP Game, but their model makes many ESP game-specific assumptions and restrictions. In contrast, the output agreement class defined here covers a far broader setting than image labeling and we do not make any assumptions or restrictions on player strategies.

Setting

Here, we formally define mechanisms for information elicitation without verification (IEWV). In the IEWV setting, there is a set of players, each holding some private information. A mechanism designer queries each player separately and simultaneously (*i.e.*, without communication between players). The designer selects an outcome of the mechanism and assigns monetary payments to each agent. Thus the mechanism, when applied to particular players, induces a Bayesian simultaneous-move game.

Player Information

To model incomplete information, we adopt the general *states of the world* model, which has been widely used in economics for modeling private information (Aumann 1976; McKelvey and Page 1986; Nielsen et al. 1990; Ostrovsky 2012). There is a finite set of possible states of the world Ω, shared by all players. An *event* is a subset of Ω; for example, the event $Q \subseteq \Omega$ could be "it is raining outside" and would consist of every state of the world in which it is raining. Nature selects a true state of the world $\omega^* \in \Omega$; an event Q is said to *occur* if $\omega^* \in Q$. Thus, the true state of the world implicitly specifies all events that occur or do not: whether it is raining, whether Alice speaks French, whether $\mathsf{P} = \mathsf{NP}$, and so on.

A player's knowledge is specified by a *prior distribution* $\mathcal{P}[\omega]$ on Ω along with a partition Π_i of Ω. A *partition* of a set Ω is a set of nonempty subsets of Ω such that every element of Ω is contained in exactly one subset. When the true state of the world is ω^*, each player i

learns the element of their partition that contains ω^*, denoted $\Pi_i(\omega^*)$. Informally, i knows that the true state of the world ω^* lies somewhere in the set $\Pi_i(\omega^*)$, but is unsure where; more precisely, i updates to a *posterior distribution* $\Pr[\omega \mid \Pi_i(\omega^*)] = \Pr[\{\omega\} \cap \Pi_i(\omega^*)] / \Pr[\Pi_i(\omega^*)]$. In line with literature on information elicitation, $\Pi_i(\omega^*)$ will be referred to as i's *signal*. (In mechanism design terms, it is player i's type.)

Throughout, we let the the set of states Ω and the number of players $n \geq 2$ be fixed.

A particular set of n players is therefore modeled by an *information structure* $\mathcal{I} = (\mathcal{P}[\omega], \Pi_1, \ldots, \Pi_n)$, where each Π_i is a partition for player i and all players share the prior $\mathcal{P}[\omega]$. $\mathcal{I}$ is common knowledge; this is the standard Bayesian game setting. We use $\mathtt{I}$ to denote the set of valid information structures on Ω with n players.

Common knowledge. Using partitions of the state space to model private information allows an intuitive formal definition of common knowledge.[2] Given partitions $\{\Pi_1, \ldots, \Pi_n\}$, the *common-knowledge partition* Π is defined to be the meet of these partitions. The *meet* of a set of partitions of Ω is the finest partition of Ω that is coarser than each individual partition. Partition Ψ is *coarser* than partition Γ (or is a *coarsening* of Γ) if each element of Ψ is partitioned by a subset of Γ. In this case, Γ is *finer* than Ψ (or is a *refinement* of Ψ).

Intuitively, an event (set of states) is *common knowledge* if, when the event occurs, all players always know that the event occurred; all players know that all players know this; and so on. The common-knowledge partition consists of the minimal (most specific) common-knowledge events.

To illustrate the difference between prior beliefs, common knowledge, and a player's posterior or private information, consider the example of labeling images. We may formalize the set of states of the world as a a list of binary attributes describing the image in full detail: "(is a dog, is not brown, is not candy, has grass in background, is running, is not a dachshund, ...)". In this case, a player's partition indicates which attributes she can distinguish; for instance, "is a dog" or not, "is a dachshund" or not, etc.

In this case, the prior is a distribution on all possible lists of attributes that an image might have. Then, once the player sees an image, she updates to a posterior. She will know several attributes for certain due to her partition; and for those that she is unsure of, she will have a posterior on them according to a Bayesian update.

The common knowledge between players in this case is the set of attributes that both players always observe. For instance, if both players can distinguish dogs from non-dogs, then whether the image is a dog will be common knowledge. But if one player cannot distinguish dachshunds from non-dachshunds, then whether the image is a dachshund will not be common knowledge.

[2] Another common approach to modeling private information is the "signals" model in which nature selects some hidden event and there is a common prior over the joint distribution of players' signals conditional on the event. This model is used in peer prediction, for example. The two models are equivalent in that each can model any scenario described by the other.

Mechanisms, Games, and Equilibria

A *mechanism* for IEWV consists of, for each player i, a report space A_i and a reward function $h_i : \mathtt{I} \times A_1 \times \cdots \times A_n \to \mathbb{R}$ that takes the player reports and returns the reward for player i (and may depend on the information structure).

When a particular group of players participate in a mechanism M, we have a Bayesian simultaneous-move game, defined as $G = (M, \mathcal{I})$. Nature selects a state of the world ω^*, each player i observes $\Pi_i(\omega^*)$ and updates to a posterior according to the prior, each i makes a report $a_i \in A_i$, and each is paid according to h_i.

A *strategy* for player i is a function s_i that specifies, for each element $\Pi_i(\omega)$ of i's partition, a probability distribution on A_i. In state ω^*, i learns element $\Pi_i(\omega^*)$ of his partition and draws an action $a_i \sim s_i(\Pi_i(\omega^*))$. A strategy profile $(s_1, \ldots, s_n)$ is a *Bayes-Nash Equilibrium* (or just *equilibrium*) of the game G if every player's strategy s_i is a *best response* to s_{-i} (the profile with s_i omitted): For every state of the world ω^*, the probability distribution $s_i(\Pi_i(\omega^*))$ on A_i is an optimal solution to

$$\max_{s_i'(\Pi_i(\omega^*))} \sum_{\omega \in \Pi_i(\omega^*)} \Pr[\omega \mid \Pi_i(\omega^*)] \, \mathbf{E}_\omega(s_i'),$$

with

$$\mathbf{E}_\omega(s_i') = \mathbb{E}\left[h_i^M(\mathcal{I}, s_1(\Pi_1(\omega)), \ldots, s_i'(\Pi_i(\omega^*)), \ldots, s_n(\Pi_n(\omega)))\right],$$

where the expectation is taken over the actions a_j drawn from each $s_j(\Pi_j(\omega))$, $j \neq i$, and a_i drawn from $s_i'(\Pi_i(\omega^*))$. The strategy profile $(s_1, \ldots, s_n)$ is a *strict* equilibrium if every s_i is the unique best response to s_{-i}.

It is most common in the literature for IEWV to construct mechanisms where the "good" (usually meaning "truthful") equilibrium is *strict*. We also wish to design focus on strict equilibria for both theoretical and pragmatic reasons.

First, in human computation mechanisms, computing and reporting a truthful response may not be the easiest or most natural strategy. For instance, on a multiple choice questionnaire, simply selecting *(a)* for every answer may be easier than picking a truthful response, if rewards are equal. So it is not clear that agents will prefer truthful reporting. Second, such mechanisms are often operated in noisy environments such as Mechanical Turk; strict incentives may encourage more accurate and less noisy responses. Finally, if one does *not* desire strict incentives, there is a natural mechanism: Ask players to report truthfully and pay them a constant amount. So, usually, the case where strict incentives are desired is more interesting from a theoretical perspective.

Queries and Specificity

We introduce the notion of a *query* associated with a mechanism. For motivation, consider the example of eliciting a prediction for the total snowfall in a city during the following year. A player's signal could be very complex and include observations of many meterological phenomena. Yet, the designer does not wish to elicit all of this weather data, only to

know a single number (predicted meters of snowfall). Thus, the designer would like to ask players to map their knowledge into a report of a single number. This mapping — from weather knowledge to predicted snowfall — is the "query" of the mechanism.

Formally, a query $T = (T_1, \ldots, T_n)$ specifies, for each player i, a function $T_i : \Delta_\Omega \to A_i$ mapping a posterior distribution to the "correct"' report when the player has that posterior belief.[3]

For example, the query could be to report the posterior distribution itself, or the expected value of some random variable, or the set of states on which the posterior has positive probability (that is, i's signal).

In mechanism design, we usually focus on *direct-revelation* mechanisms where players are simply asked to report their signal. However, in IEWV, it is of interest to consider other queries as well. One reason for this is that we are interested in descriptively modeling non-direct-revelation mechanisms, like output agreement, that exist in the literature or in practice. A second reason to consider general queries is because this makes our impossibility results stronger — they apply to mechanisms attempting to elicit *any* type of information.

Specificity. Here, we generalize truthfulness to *specificity* of player reports, capturing the following question: *What knowledge does a player use in reporting an answer to a query?* To our knowledge, this work is the first to consider such an extension to the traditional notion of truthfulness.

Given a query T and a partition $\hat{\Pi}$, define the notation $T_{\hat{\Pi}}$ to be the strategy that, for each ω^* chosen by nature, makes the report $T(\Pr[\omega \mid \hat{\Pi}(\omega^*)])$. In other words, $T_{\hat{\Pi}}$ reports correctly according to the posterior distribution induced by $\hat{\Pi}$. Notice that a player i can only play strategy $T_{\hat{\Pi}}$ if $\hat{\Pi}$ is a coarsening of his partition Π_i: Otherwise, he will not in general know which element of $\hat{\Pi}$ contains ω^*.

Definition 1. *A player i's strategy s_i is called* $\hat{\Pi}$-specific *if:*

1. $\hat{\Pi}$ *is a coarsening of i's partition* Π_i*, and*
2. $s_i = T_{\hat{\Pi}}$.

To gain intuition, we note three natural special cases. The case $s_i = T_{\Pi_i}$, or Π_i-specificity, is just truthfulness: always reporting according to i's posterior. On the other extreme, the case $s_i = T_{\{\Omega\}}$, or $\{\Omega\}$-specific, means always reporting according to the prior no matter what signal is received. In the middle, we identify the case $s_i = T_{\Pi}$, or Π-specific, or *common-knowledge specific*: reporting according to common knowledge.

Any strategy that is $\hat{\Pi}$-specific, for some coarsening $\hat{\Pi}$ of their partition Π_i, has two nice properties that one might associate with "weak" truthfulness. We illustrate with a running example: Suppose a player observes today's date, and consider coarsenings $\hat{\Pi}_1 =$ the twelve months of the year and $\hat{\Pi}_2 =$ the four seasons. First, specificity requires that a player report according to an event that *actually occurs*. For example, given that it is August 2nd, a player may report "it is August" as with $\hat{\Pi}_1$, or "it is summer" as with $\hat{\Pi}_2$, but there is no partition where he may report that it is January or that it is spring. Second, reports must be consistent across each element of $\hat{\Pi}$. For example, if a player reports "it is summer" when it is August 2nd, then the player must make this exact same report on every other day of summer. He cannot report "it is summer" on August 2nd but report "it is August" on August 3rd.

Meanwhile, $\hat{\Pi}$ specifies the *granularity* of the information. For example, we could have month-specific or season-specific information. We thus get a partial ordering or hierarchy of specificity, with truthfulness as the best and reporting the prior as the worst, where $\hat{\Pi}_1$-specific is better than $\hat{\Pi}_2$-specific if $\hat{\Pi}_1$ is a finer partition than $\hat{\Pi}_2$.

We can now utilize specificity in defining our equilibrium solution concept: An equilibrium $(s_1, \ldots, s_n)$ is $(\hat{\Pi}_1, \ldots, \hat{\Pi}_n)$-specific if each player i plays a $\hat{\Pi}_i$-specific strategy in it; as important special cases, we identify truthful and common-knowledge-specific equilibria.

Equilibrium Results

Here, we provide a formal definition and game-theoretic analysis of the two-player output agreement class of mechanisms. We show that the mechanisms elicit *common-knowledge-specific* reports with strict incentives. We then show that this is the best that can be hoped for by any mechanism making as few assumptions on the information structure as output agreement; we also show that the existence of uninformative equilibria is unavoidable.

Definition 2. *A two-player* output agreement mechanism M *is a mechanism for eliciting information without verification defined as follows. The mechanism designer announces a report space $A = A_1 = A_2$ and an associated query T where $T_1 = T_2$ (we will abuse notation by just writing T rather than T_i). The designer selects a distance metric d on the space A and a monotonically decreasing reward function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}$. Each player i makes a report $a_i \in A$ and is paid $h_i^M(a_1, a_2) = h(d(a_1, a_2))$.*

A *distance metric* $d : A \times A \to \mathbb{R}$ satisfies that $d(x, y) \geq 0$ with equality if and only if $x = y$, that $d(x, y) = d(y, x)$ for all $x, y \in A$, and that $d(x, y) \leq d(x, z) + d(y, z)$ for all $x, y, z \in A$.

For an example mechanism in this category, consider an audio transcription task: Two players each listen to a thirty-second clip of speech and are asked to produce the written transcription. The distance function on their outputs (transcripts) is Levenshtein (edit) distance. The reward function can be a fixed constant minus the edit distance between their transcripts.

[3] One could generalize in two ways: First, by allowing multiple possible correct answers for a given posterior, so that T_i maps to a *set* of responses; and second, by allowing queries to specify *randomized* reports, where the player is asked to draw from some distribution. Output agreement can be generalized to include such cases, although the notion of strict equilibrium requires tweaking; and similarly, our negative results extend to these cases as well even for "tweaked" equilibrium concepts.

Theorem 1. *For any query T, any output agreement mechanism with a strictly decreasing reward function elicits a strict equilibrium that is common-knowledge-specific for T.*

Proof. For each player i, let s_i be a Π-specific strategy with respect to T; that is, $s_i(\Pi_i(\omega^*)) = T(\Pr[\omega \mid \Pi(\omega^*)])$.

Since Π is the common knowledge partition, we have that in every state ω^*, $s_1(\Pi_1(\omega^*)) = s_2(\Pi_2(\omega^*))$. In any state, both players' strategies put full support on the same report; thus, each player does strictly worse by drawing from any other distribution. Thus (s_1, s_2) is a strict equilibrium. □

How positive is Theorem 1, and can it be improved upon? It is quite positive along the "query" axis: It works for *any* given query. Prior mechanisms for IEWV tend to focus primarily on eliciting signals. However, along the "specificity" axis, we might naively hope for better; for instance, we might want a *truthful* mechanism. But, notice that output agreement makes no assumptions on the information structure $\mathcal{I}$ of the players. In the next section, we show that *no* mechanism can strictly improve on common-knowledge specificity unless it makes some such assumption. This shows that output agreement is actually optimal along the specificity axis among the class of mechanisms that make no assumptions on $\mathcal{I}$.

Impossibility Results

In this section, we give two broad impossibility results for IEWV. First, as just discussed, we show that no mechanism can guarantee an equilibrium more specific than common knowledge unless it makes some assumption on the information structures.

Second, we address a different concern about output agreement mechanisms, that they have "bad" equilibria: Players can agree beforehand to all make the same report, ignoring their signals. Our second impossibility result says that the same is true of all mechanisms for IEWV.

Theorem 2. *Let T be any query and M any mechanism for IEWV. Then M cannot guarantee a strict equilibrium more specific than common knowledge. In particular, there is some information structure $\mathcal{I}$ for which M is not strictly truthful.*

The proof creates an information structure where one player's partition is finer than the other's, then shows that the other player (and thus the mechanism's reward rule) cannot distinguish between two different posteriors of the first player.

Uninformative equilibria. In IEWV, the goal is to design mechanisms with "good" equilibria in which information is revealed. However, it has previously been noted informally and observed for individual mechanisms or special cases (Lambert and Shoham 2008; Jurca and Faltings 2005; Della Penna and Reid 2012) that such mechanisms often also have equilibria that are "bad" in some way. The conjecture that this holds more generally may be considered something of a suspected folk theorem in the literature.

The following characterization formalizes this intuition in a very broad setting and for very "bad" equilibria: those in which absolutely no information is revealed. Intuitively, the characterization says that, if we take a game of IEWV and ignore the signals received by each player, we can treat it as a game of *complete information* (*e.g.* in normal form); under very weak conditions, this game has an equilibrium, and we can show that this equilibrium is an "uninformative" equilibrium in the original game of IEWV.

Theorem 3. *A strategy is termed* uninformative *if it draws actions from the same distribution in every state of the world (*i.e. *for every signal observed). A game of IEWV has an equilibrium made up of uninformative strategies if and only if there exists a Nash equilibrium in the two-player complete-information game whose payoff matrix is given by its reward rule.*

Player Inference and Focal Equilibria

Suppose that, in an output agreement game, player 1 is presented with a given query; she might initially consider a Π_1-specific (truthful) strategy. But she knows that player 2 should play a best response, which in general is not necessarily Π_2-specific; and then she (player 1) should switch to a best response to *that* strategy, and so on. We refer to this the process of computing a sequence of best response strategies as *player inference*. [4] An example player inference process is given in Figure 1; it gives an example where players are asked to report the most likely realization of a random variable, which may be either ★ or △. We revisit the example in Theorem 5.

Ideally, this inference process would converge to the common-knowledge-specific equilibrium (since it was shown in the previous section that this is the "best" equilibrium). We can show that this does indeed happen when eliciting the mean of a random variable.

Theorem 4. *Let t be a random variable taking values in $\mathbb{R}^n$. There is an output agreement mechanism for eliciting the mean of t such that any sequence of best response strategies, beginning with a Π_i-specific strategy, converges to a Π-specific equilibrium.*

The proof is somewhat notationally involved and utilizes a result of (Samet 1998), but is straightforward. The intuition is to reward both players by the Euclidean distance between their reports, $h(x, y) = -d(x, y)^2$ where $d(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$. With this choice, a best response is exactly the expected value of the other player's report; iterated best responses involve iterated expectations over various subsets of states, weighted by various posterior probabilities on these states; and on average, the weight on each state converges to the common-knowledge posterior probability of that state. (The heavy lifting in proving this is done by (Samet 1998).) This gives an expectation of t according to common knowledge.

This result is encouraging because many natural tasks may be modeled as reporting the mean of some random variable. These could include straightforward numerical queries such

[4] It is of note that this process does not consist of players taking or observing actions (as opposed to *best-response dynamics* and *fictitious play*); rather, it is the hypothetical process of a rational agent computing the optimal strategy to play.

$\underline{\omega_1}$	$\underline{\omega_2}$	$\underline{\omega_3}$
$\mathcal{P}[\omega_1] = 0.40$	$\mathcal{P}[\omega_2] = 0.35$	$\mathcal{P}[\omega_3] = 0.25$
$t = \bigstar$	$t = \triangle$	$t = \triangle$

(a) The three possible states of the world $\omega_1, \omega_2, \omega_3$ with their prior probabilities and the value of the random variable t in each.

$\underline{\{\omega_1\}}$	$\underline{\{\omega_2, \omega_3\}}$
$\Pr[\omega_1] = 1.0$	$\Pr[\omega_2] = 0.58, \Pr[\omega_3] = 0.42$
mode $= \bigstar$	mode $= \triangle$

(b) Player 1's signal structure: the left signal when the state is ω_1, the right when it is ω_2 or ω_3. For each signal, the posterior beliefs and the "mode" (most likely value) of t.

$\underline{\{\omega_1, \omega_2\}}$	$\underline{\{\omega_3\}}$
$\Pr[\omega_1] = 0.53, \Pr[\omega_2] = 0.47$	$\Pr[\omega_3] = 1.0$
mode $= \bigstar$	mode $= \triangle$

(c) Player 2's signal structure: the left signal when the state is ω_1 or ω_2, the right when it is ω_3; posterior beliefs and mode of t for each. For both signals observed, if player 1 is reporting truthfully, then player 2's best response is to be truthful.

$\underline{\{\omega_1\}}$	$\underline{\{\omega_2, \omega_3\}}$
$\Pr[\omega_1] = 1.0$	$\Pr[\omega_2] = 0.58, \Pr[\omega_3] = 0.42$
response $= \bigstar$	response $= \bigstar$

(d) Player 1's signals and posterior beliefs again, this time showing the best response when player 2 is reporting truthfully. Player 2's best response to this strategy will be to also always report $\bigstar$, and they will be in equilibrium.

Figure 1: Information structure for an output agreement game. Players are asked to report the "mode" (most likely value) of t, which could be either $\bigstar$ or $\triangle$. The players are paid 1 if they agree and 0 if they disagree. A player's best response given her signal is whichever of $\bigstar$ or $\triangle$ is more likely to be reported by her opponent. In this example, if we start with a truthful strategy from either player and iteratively compute best response strategies, we converge to an equilibrium where both players always report $\bigstar$ no matter what they observe. (Furthermore, it is more likely that t is actually $\triangle$.

as estimating the number of cells in a microscope image; geographical tasks such as estimating the facility location that would minimize average commute time for a large population; or numerical prediction tasks for long-term events like yearly snowfall (where waiting to reward agents until ground truth becomes available may be undesirable).

However, this nice convergence result does not extend to two of the other most natural properties: median and mode. In fact, this holds more broadly than in $\mathbb{R}^n$; we consider (non-constant) random variables taking values in an arbitrary metric space. By *median* of t, we mean a value in the range of t that minimizes the expected distance to $t[\omega]$. By *mode*, we mean a value in the range of t with highest total probability.

Theorem 5. *When $|\Omega| \geq 3$, no output agreement mechanism for eliciting the median or mode of a random variable in an arbitrary metric space ensures for all settings that a sequence of best response strategies, beginning with a Π_i-specific strategy for either player i, converges to a Π-specific equilibrium.*

The key counterexample that proves this statement is given in Figure 1. Note that in state ω_3, both players are certain that the true realization of the random variable is $\triangle$, yet both report $\bigstar$ due to their uncertainty about the other's report. Furthermore, this may be generalized to an arbitrarily bad example. Intuitively, let the true realization be $\triangle$ with probability $1 - \epsilon$, and let each player's partition divide up the state of the world into sets with probability 2ϵ, but all overlapping (so each element of 1's partition has ϵ overlap with each of two different elements of 2's partition, and vice versa). When the realization is $\bigstar$, player 1 always observes this but player 2 is unsure. Now by modifying probabilities slightly to break ties "toward" $\bigstar$, we can cause a cascading sequence of best responses so that, at the end, both players always report $\bigstar$ even though the realization is almost always $\triangle$.

Conclusions

Output agreement is a simple and intuitive mechanism. However, when formalized and examined from the point of view of information elicitation without verification, it raises surprisingly complex questions. These include the notion of *specificity* of player reports and the identification of common-knowledge-specific equilibria in output agreement, as well as the question of player inference and focal equilibria in this setting. We hope that these concepts will find use outside of output agreement mechanisms in the IEWV literature.

Output agreement mechanisms, meanwhile, are interesting in their own right, providing several advantages over other mechanisms. First, they do not require the mechanism designer to assume anything about the signal structure of the participants. Second, it is conceptually simpler and easier to explain and implement, which may be beneficial in practice. Third, it allows for any report space, which includes *e.g.* asking players to *compute* on their signals, whereas other mechanisms tend to be limited to reporting of (often binary) signals. Fourth, it is *robust* in that its equilibrium guarantee holds for *any* signal structure.

Moreover, it turns out that this last property cannot be achieved by mechanisms that elicit private information in equilibrium. Output agreement's common knowledge guarantee is the best we can hope for if we desire this robustness property. Another downside of output agreement, that it has "uninformative" equilibria, turns out to be inherent to the IEWV setting: All other mechanisms have them too. These impossibility results may also contribute to the IEWV literature by helping illustrate the nature of these difficulties.

Acknowledgements

The authors thank David Parkes for helpful discussions. This work was partially supported by the National Science Foundation, under grant CCF-0953516. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

References

Aumann, R. J. 1976. Agreeing to disagree. *Annals of Statistics* 4(6):1236–1239.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, 319–330. International World Wide Web Conferences Steering Committee.

Della Penna, N., and Reid, M. 2012. Crowd & prejudice: An impossibility theorem for crowd labelling without a gold standard. In *Collective Intelligence*, CI '12.

Goel, S.; Reeves, D.; and Pennock, D. 2009. Collective revelation: A mechanism for self-verified, weighted, and truthful predictions. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, 265–274. ACM.

Huang, S., and Fu, W. 2012. Systematic analysis of output agreement games: Effects of gaming environment, social interaction, and feedback. In *Proceedings of HCOMP 2012: The Fourth Workshop on Human Computation*.

Jain, S., and Parkes, D. 2008. A game-theoretic analysis of games with a purpose. In *Internet and Network Economics*, WINE '08, 342–350.

Jurca, R., and Faltings, B. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *Internet and Network Economics*, WINE '05, 268–277. Springer.

Jurca, R., and Faltings, B. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce*, EC '06, 190–199. ACM.

Jurca, R., and Faltings, B. 2007a. Collusion-resistant, incentive-compatible feedback payments. In *Proceedings of the 8th ACM conference on Electronic commerce*, EC '07, 200–209. ACM.

Jurca, R., and Faltings, B. 2007b. Robust incentive-compatible feedback payments. In Fasli, M., and Shehory, O., eds., *Agent-Mediated Electronic Commerce*, volume LNAI 4452, 204–218. Berlin Heidelberg: Springer-Verlag.

Jurca, R., and Faltings, B. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research* 34(1):209.

Lambert, N., and Shoham, Y. 2008. Truthful surveys. WINE '08, 154–165. Springer.

McKelvey, R. D., and Page, T. 1986. Common knowledge, consensus, and aggregate information. *Econometrica* 54(1):109–127.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51(9):1359–1373.

Nielsen, L. T.; Brandenburger, A.; Geanakoplos, J.; McKelvey, R.; and Page, T. 1990. Common knowledge of an aggregate of expectations. *Econometrica* 58(5):1235–1238.

Ostrovsky, M. 2012. Information aggregation in dynamic markets with strategic traders. *Econometrica* 80(6):2595–2647.

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 306(5695):462–466.

Radanovic, G., and Faltings, B. 2013. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, AAAI '13.

Samet, D. 1998. Iterated expectations and common priors. *Games and economic Behavior* 24(1-2):131–141.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems*, CHI '04, 319–326. ACM.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51(8):58–67.

Weber, I.; Robertson, S.; and Vojnovic, M. 2008. Rethinking the esp game. In *Proceedings of the 27th International Conference on Human factors in Computing Systems*, volume 9 of *CHI '08*, 3937–3942.

Witkowski, J., and Parkes, D. 2012a. A robust bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI '12.

Witkowski, J., and Parkes, D. 2012b. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, 964–981. ACM.

Witkowski, J.; Bachrach, Y.; Key, P.; and Parkes, D. C. 2013. Dwelling on the negative: Incentivizing effort in peer prediction. In *First AAAI Conference on Human Computation and Crowdsourcing*.

Proofs for Equilibrium Results

Theorem (Theorem 2)**.** *Let T be any query and M any mechanism for IEWV. Then M cannot guarantee a strict equilibrium more specific than common knowledge. In particular, there is some information structure $\mathcal{I}$ for which M is not strictly truthful.*

Proof. The approach will be to start by ruling out strict truthfulness, then extend to any solution more specific than common knowledge. We will consider a player whose truthful response depends on which signal she receives, but who is much better informed than her opponents. There will be two signals where her query specifies two different truthful responses, but her best responses will be the same. In any equilibrium, in one of these cases she has a best response that is not a truthful response to the query.

We need the basic assumption that the query is *nontrivial*. A query T is considered trivial if, for every $\mathcal{I}$, for each player i, $T_i(p)$ is the same for all possible posteriors $p = \Pr[\omega \mid \Pi_i(\omega^*)]$. A trivial query would mean that all players should always report the same thing no matter what information they observe.

So consider, by nontriviality, a prior $\mathcal{P}[\omega]$, player i, and partition Π_i such that, for any probability distribution p_i on A_i, there is some state in which p_i is not truthful, *i.e.* $p_i \neq T_i(\Pr[\omega \mid \Pi_i(\omega^*)])$ for some ω^*. Consider a game with prior $\mathcal{P}[\omega]$ in which player i has partition Π_i and all other players j have a trivial partition $\Pi_j = \{\Omega\}$.

Let $(s_1, \ldots, s_n)$ be a truthful equilibrium. Pick a particular state ω^*; in this state, player i plays according to the distribution $p_i^* = s_i(\Pi_i(\omega^*))$. Since $(s_1, \ldots, s_n)$ is an equilibrium, p_i^* maximizes expected utility against s_{-i} in state ω^*. But s_{-i} is constant on all states of the world (since players $2, \ldots, n$ receive the same signal in every state). So construct the strategy s_i' where, for every ω, $s_i'(\Pi_i(\omega)) = p_i^*$. We immediately have that s_i' is also a best response to s_{-i}.

But by nontriviality, there is some state $\omega' \in \Omega$ such that $p_i^* \neq T_i(\Pr[\omega \mid \Pi_i(\omega')])$. Thus, s_i' is not a truthful strategy; hence $(s_1, \ldots, s_n)$ is not a strictly truthful equilibrium.

Now, we simply notice that the proof works, not just for truthfulness, but for any $\hat{\Pi}$-specific equilibrium where $\hat{\Pi}$ is a strictly finer partition that the common-knowledge partition Π. We can construct the same counterexample in this case. □

To prove Theorem 3, we define our terms more formally and introduce the notation G' for a complete-information version of the game G.

Definition 3. *A strategy s_i for player i is* uninformative *if for all ω, ω', $s_i(\Pi_i(\omega)) = s_i(\Pi_i(\omega'))$. An equilibrium $(s_1, \ldots, s_n)$ is* uninformative *if s_i is uninformative for all i.*

Definition 4. *(G') For any Bayesian game $G = (M, \mathcal{I})$ for information elicitation without verification, let G' denote the induced simultaneous-move game of complete information where each player i selects and reports an action $a_i \in A_i$ and receives a payoff of $h_i^M(\mathcal{I}, a_1, \ldots, a_n)$. A strategy in G' is a probability distribution over actions; a profile of best response strategies is a Nash equilibrium.*

Theorem (Theorem 3 restated)**.** *A game G of information elicitation without verification has an uninformative equilibrium if and only if there exists a Nash equilibrium in G'.*

Proof. We show a one-to-one correspondence between the two. First, we note that strategy sets in G' are vectors of probability distributions $(p_1, \ldots, p_n)$ from which players draw their actions. Second, we note that uninformative strategy sets in G are determined uniquely by a vector of distributions $(p_1, \ldots, p_n)$, because for each i and for all $\omega, \omega' \in \Omega$, $s_i(\Pi_i(\omega)) = s_i(\Pi_i(\omega')) = p_i$. Therefore, there is a one-to-one correspondence between strategy sets in G' and uninformative strategy sets in G. But each player i's reward for a realized profile of actions $(a_1, \ldots, a_n)$ is identical in G' and in G (by construction of G'). So when each player j draws an action from p_j, drawing actions from to p_i maximizes i's expected utility in G' if and only if it does so in G. This completes the proof. □

Proofs for Player Inference Results

Theorem (Theorem 4)**.** *Let t be a random variable taking values in $\mathbb{R}^n$. There is an output agreement mechanism for eliciting the mean of t such that any sequence of best response strategies, beginning with a Π_i-specific strategy, converges to a Π-specific equilibrium.*

In our context, a *random variable* taking values in some space X is a mapping $t : \Omega \to X$ where $t[\omega]$ specifies the value of t when the state of the world is ω. Thus the query for eliciting the mean in $\mathbb{R}^n$ is $T(p(\omega)) = \mathbb{E}_{\omega \sim p}\, t[\omega]$.

Proof. We select d to be the Euclidean distance and h to be any affine transformation of $-x^2$. This choice ensures that a player's best response is to report her expected value of her opponent's report. More formally, it is straightforward to verify that the unique maximizer of $\mathbb{E}_\omega\, h(d(a, t[\omega]))$ is $a = \mathbb{E}_\omega\, t[\omega]$, where the expectation is taken according to the same distribution in both cases.

We first note that a Π_i-specific report of the mean puts full support on a single response; likewise, by the above, all best responses put full support on a single response (since each is an expected value). Therefore, when considering a sequence of best response strategies beginning with a Π_i-specific one, we need only consider such strategies.

Now we will view a player's strategy s_i as a random variable F_i where $F_i[\omega]$ is the report given full support by $s_i(\Pi_i(\omega))$. Consider the sequence $t, F_i^{(1)}, F_j^{(2)}, F_i^{(3)}, F_j^{(4)}, \ldots$; this is in correspondence with a sequence of best response strategies where $F_i^{(1)}$ is Π_i-specific and each random variable in the sequence consists of expectations of the previous variable according to the appropriate posterior beliefs. Formally, in each state ω^*, $F_j^{(k)}[\omega^*] = \sum_\omega \Pr[\omega \mid \Pi_j(\omega^*)]\, F_i^{(k-1)}[\omega]$, and the same holds with i and j reversed. This construction allows us to use the following nice result of Samet:

Lemma 1 (Theorem $2'$ and Theorem $1'$ of (Samet 1998))**.** *Let t be a random variable taking values in $\mathbb{R}$ and consider the sequence $t, F_i^{(1)}, F_j^{(2)}, \ldots$ of iterated expected values restricted to states of a fixed element Q of the common-knowledge partition Π. If and only if player beliefs are consistent with the existence of a common prior, then this sequence converges on states in Q, and its value in each state $\omega^* \in Q$ is the same; moreover, this value is $\sum_\omega \Pr[\omega \mid Q]\, t[\omega]$.*

This gives that, when $t \in \mathbb{R}$, the sequence of iterated expected values converges to the common-knowledge expected value. To use this result, consider any fixed $Q \in \Pi$. We note that the expected value of a random variable in $\mathbb{R}^n$ is an n-tuple whose k-th entry is the expected value of the k-th entry of the random variable. Therefore, for each $k = 1, \ldots, n$, a sequence of best responses $F_i^{(1)}, F_j^{(2)}, \ldots$ involves alternately computing, for each ω^*, the expected value of the previous strategy's k-th entry. Therefore, by Samet, the k-th entry of the best response converges to the expected value of the kth entry of t according to the common-knowledge posterior when $\omega^* \in Q$.

Because this holds for all entries k of t, this implies that in every state ω^*, the player strategies converge to reporting the expected value of t according to the common-knowledge element $\Pi(\omega^*)$. Finally, by Theorem 1, we have that reporting

the common-knowledge mean actually is an equilibrium. So the inference process converges to the equilibrium where players report the common-knowledge mean. □

Theorem (Theorem 5)**.** *When $|\Omega| \geq 3$, no output agreement mechanism for eliciting the median or mode of a random variable in an arbitrary metric space ensures for all settings that a sequence of best response strategies, beginning with a Π_i-specific strategy for either player i, converges to a Π-specific equilibrium.*

Proof. We first demonstrate that a necessary condition for a sequence of best response strategies to converge to a Π-specific equilibrium would be that the composition of reward function h and distance metric d be a *strictly proper scoring rule*[5] for the given property (median or mode). We then show that no mechanism with this property is successful by constructing a counterexample for any nontrivial report space.

Consider the case where player 2 is completely informed: the partition Π_2 consists of singleton sets. In every state of the world, player 2 learns the exact value of t. Let player 1 have some strictly coarser partition. Now consider a Π_2-specific strategy of player 2; player 1 has some best response s_1. Then player 2's best response will be to exactly mimic this strategy; player 2 can set $s_2(\{\omega^*\}) = s_1(\Pi_1(\omega^*))$ for every state ω^*. Both players receive the maximum reward in every state, so this is an equilibrium.

This equilibrium is common-knowledge-specific whenever player 1's best response to a Π_2-specific strategy is Π_1-specific (since $\Pi_1 = \Pi$ in this case). But for either median or mode, a Π_2-specific strategy is to simply report the value of the random variable in the observed state. So player 1 faces a scoring rule $h(d(s_1(\Pi_1(\omega^*)), t[\omega^*]))$ in state ω^*. Player 1 has a strict incentive to best-respond truthfully according to Π_1 if and only if h is a strictly proper scoring rule for the appropriate property.

Thus, to elicit either the median or mode, h composed with d must be a strictly proper scoring rule; that is, a best response must be to report the median (respectively, mode) of an opponent's strategy. Now construct a counterexample of a random variable whose range consists of two distinct values. Without loss of generality, suppose there are only three states of the world (otherwise, split them into three groups arbitrarily and treat each group as a state). Then we construct the information structure in Figure 1, where ★ and △ represent the two values in the range (and with an arbitrary, but fixed, distance between them). In this case, the median and mode necessarily coincide, and a best response (as argued above) must be the mode of the opponent's strategy. As demonstrated in Figure 1, the common-knowledge most likely value is △; however, the inference process always reaches an equilibrium in which ★ is always reported. □

[5]In this context, a *scoring rule* $S(a, a^*)$ takes a report $a \in A$ and a realization $a^* \in A$ of a random variable and returns a payoff; it is *strictly proper* for a property if, for fixed beliefs, reporting the value of the property according to those beliefs uniquely maximizes expected score according to those beliefs.