Aumann Agreement Game

Game by Scott Garrabrant
Written by Abram Demski

Aumann's Agreement Theorem states that two Bayesians with common knowledge of each other's posteriors cannot agree to disagree: taking each other's disagreement as evidence, they must come to an agreement. The theorem assumes not only perfect rationality, but perfect trust in each other's rationality, trust in that trust, and so on! This game is about seeing how things play out given the *imperfect* trust between you and your friends. It's a multiplayer version of the *calibration game*; those who have not heard of that may want to look it up and play that first. It's not required, however; this game teaches the same lessons (and more!).

Setup

1. Find a suitable set of multiple-choice trivia questions. The number of players equals the number of answers; we give a scoring for 4 players. In addition, a moderator is needed to read the questions.
2. Each player needs paper on which to record their score. The moderator needs small slips of paper to write on. It's also good for the moderator to have a small whiteboard.

Game Play

1. For each question, the moderator must write down the possible answers and distribute one to each player. The question is read aloud, but the possible answers are kept secret: each player knows only the one they've been given.
2. Each player declares a probability estimate for their answer being the correct answer whenever they are ready, and announce revisions as many times as they want. It can be fun for the moderator to write down each announced probability in sequence (we use a small whiteboard for this). Only the final probabilities matter for score, however. Note: the probabilities the players settle on need not sum to 1. Players might want to adjust toward summing to 1, but not if they don't trust each other's calibration enough!
3. When no more revisions are announced, the moderator asks if everyone is satisfied with their final probability estimate. If so, the correct answer is revealed, and players reveal the answers they each held.
4. The players are scored based on their final probabilities, and whether the answer they held was correct or incorrect. For four players, score as follows. For final probability p:

    Holding the *correct* answer:

    $100*\log_2(4p)$

Holding an *incorrect* answer:

$100*\log_2((4/3)*(1-p))$

A table for this score is given later in this document. If a player gives probability 0 and it turns out they're holding the correct answer, or probability 1 and it turns out they're holding the incorrect answer, their score is negative infinity and they cannot recover.

5. Each round proceeds with a new question. After a few rounds, consider scoring the players and switching moderators.

Example Round

Abram, David, Harris, and Scott are players, with Joanna moderating. The question is "What is the capital city of Estonia?" Joanna writes down the four possible answers and distributes them to the players:

Abram - "Tartu"
David - "Tallinn"
Harris - "Warsaw"
Scott - "Lodz"

Abram is completely uncertain whether his slip of paper holds the correct answer, and announces a probability of 25%. Scott joins Abram at 25%. David reports a 90% probability, and Harris 33%. Abram thinks David probably has the correct answer and Harris might if not David, so revises down to 5%. Scott thinks David may have the right answer but is probably overconfident and revises down to 10%, ignoring Harris because he's never sure what Harris is up to anyway. Harris revises up to 95% to see if he can get a reaction. David drops his probability to 50% in response to this. Harris backs down to 5% because really he's fairly sure he doesn't have the answer. David throws his hands in the air and goes back up, but only to 85% because Harris made the possibility of being wrong more vivid. Abram and Scott stick with their answers. Joanna wrote the following on the whiteboard:

Abram	.25					.05								
David			.90							.50			.85	
Harris					.33				.95			.05		
Scott		.25					.10							

Joanna confirms the final probabilities, and announces that the correct answer is Tallinn. David, the holder of this answer, gets a score of 177 for this round. Abram gets 34, Harris gets 34, and Scott gets 26. David curses his uncertainty (and maybe Harris); he almost had 185 points. Abram wipes the sweat off his brow, feeling in hindsight that he stuck his neck out saying .05 and got lucky. If his answer had turned out to be the correct one after all, he would have lost 232 points.

Scoring Table

%	Cor	Inc	%	Cor	Inc	%	Cor	Inc	%	Cor	Inc	%	Cor	Inc
1	-464	40	21	-25	7	41	71	-35	61	129	-94	81	170	-198
2	-364	39	22	-18	6	42	75	-37	62	131	-98	82	171	-206
3	-306	37	23	-12	4	43	78	-40	63	133	-102	83	173	-214
4	-264	36	24	-6	2	44	82	-42	64	136	-106	84	175	-223
5	-232	34	25	0	0	45	85	-45	65	138	-110	85	177	-232
6	-206	33	26	6	-2	46	88	-47	66	140	-114	86	178	-242
7	-184	31	27	11	-4	47	91	-50	67	142	-118	87	180	-253
8	-164	29	28	16	-6	48	94	-53	68	144	-123	88	182	-264
9	-147	28	29	21	-8	49	97	-56	69	146	-127	89	183	-277
10	-132	26	30	26	-10	50	100	-58	70	149	-132	90	185	-291
11	-118	25	31	31	-12	51	103	-61	71	151	-137	91	186	-306
12	-106	23	32	36	-14	52	106	-64	72	153	-142	92	188	-323
13	-94	21	33	40	-16	53	108	-67	73	155	-147	93	190	-342
14	-84	20	34	44	-18	54	111	-71	74	157	-153	94	191	-364
15	-74	18	35	49	-21	55	114	-74	75	158	-158	95	193	-391
16	-64	16	36	53	-23	56	116	-77	76	160	-164	96	194	-423
17	-56	15	37	57	-25	57	119	-80	77	162	-171	97	196	-464
18	-47	13	38	60	-27	58	121	-84	78	164	-177	98	197	-523
19	-40	11	39	64	-30	59	124	-87	79	166	-184	99	199	-623
20	-32	9	40	68	-32	60	126	-91	80	168	-191			

Scoring Explanation

The scores given are adjusted Bayes scores. The Bayes score is a *proper scoring rule:* a way of assigning points so that the optimal strategy is to report honest probabilities.

The raw Bayes score is the logarithm of the probability you assign to the event that happened. This is somewhat difficult to interpret. To make it more meaningful, we can compare to the Bayes score of the maximum-uncertainty (maximum-entropy). In the case of a 4-player game, this means giving probability .25 no matter what. If the player holds the correct answer, this achieves a score of .25 on the round; if they hold an incorrect answer, .75. Dividing by these numbers gives the $\log_2(4p)$ and $\log_2((4/3)*(1-p))$ in the scoring formula. The adjusted score we compute will be positive when we do better than this strategy, and negative when we do worse. We also multiply the score by 100 to get integers. The adjusted score also has the benefit of adjusting for good and bad luck in the answers players are dealt: otherwise, a player who is dealt incorrect answers repeatedly will tend to have a better score. (It's easier to guess that you're answer isn't correct than to guess that it is.)