Article on Anti-Cheating Online

Since competitive over-the-board chess in England ceased near the end of March 2020, ECF and 4NCL has done very well in setting up all sorts of online competitions – an online 4NCL and junior equivalent, an English Online Blitz Championship, a County Championship and Club Championship; and an English Online Youth Championship. There are regular weekly internal events for England Juniors and England Women, and now the UK Chess Challenge is running online. English chess in general is far more devolved in terms of its administration than a number of other Federations around the world, and in these times it meant that a number of people have got on with a number of projects without needing to lean on the ECF to make any of them happen. We can be very proud of what we've achieved.

Nevertheless, as thousands of English over-the-board players have suddenly landed on various online platforms and got acquainted with something in which they had minimal interest even four months ago, organisers have had to learn entirely new skills that they hadn't previously needed to quite the same extent. At a trivial level, this meant adapting competitions to have formats that can cater with the "off-the-shelf" solutions provided by the websites rather than a blank sheet of paper. The most complex of all the issues is, of course, the issue of fair play.

It is important at this stage to make it clear that I do not work for any of the platforms! I have been working on a number of the aforementioned events predominantly played in by English players, and so much of what I am writing is based purely on my experience with those. I have no vested interest in saying certain platforms are good and certain players are not; the purpose of this is to set out what I have learnt for the benefit of those who might come later.

The 4NCL has a two-way process towards handling these matters, which we believe has served us well:

- Lichess has its own anti-cheating systems. They don't disclose how they work on the basis that if they did, people who are determined to cheat will uset that information get around it. In general, they have two levels of protection: so obvious that their computer detection calls it automatically, and a level requiring the intervention of a moderator for cases that either the computer detection thinks is questionable, or one or more other website users have reported to them as questionable. This general approach appears to be how all websites handle this. It is important to stress that the overwhelming majority of users on a chess platform are unknown to them, so their systems are detecting computer assistance rather than an improvement relative to their play.

- We use Professor Ken Regan's model, which has existed in various forms since the bitter Topalov-Kramnik World Championship Match and has been used in FIDE and national cases since 2011. To some extent this is much more simplistic than what the websites have, because all that is available are the PGNs of the game – i.e. the moves that were made. Professor Regan will therefore always get far less clear results than a website will due to having fewer inputs, but it is useful in three ways: it has a rapid first step that filters down the people who need to be reported to the platform (the "two-way process" I referred to earlier), its second stage does predictive analytics and factors in the standard rating of the player, which the platorm can't in the majority of cases because the player is unknown to them – the 4NCL knows who all of the players and their over-the-board ratings, and finally, the results from Ken's data can be released to the player if they have been deemed to have used computer assistance, because doing so would not compromise his model.

Both of these systems have one fundamental question at the heart of them: How can you use any data to tell whether or not someone is using computer assistance?

Step 1 of Professor Regan's system looks for two characteristics about a given move:

- Move Match Percentage. Does the player play the engine's first choice?
- Average Scaled Difference. This is the average error per move judged by computer but scaled using a method published in academic papers with Guy Haworth in 2011, so that differences in

uneven positions count less. We don't know, but we think that this is very similar to the "average centipawn loss" that Lichess provides if you get the computer to analyse your game afterwards.

Step 2 uses deep analysis of all reasonable moves and treats the opening book in detail, whereas Step 1 is just a screening designed to give advice to cases where more detailed information is needed. There are two levels of depth in stage 1, a lead of 4-pawns or a lead of 9-pawns – 9-pawns is really only used for Rapid and Blitz.

Step 1 will provide a ranking list of who has the biggest match on these two characteristics, but it is insufficient to say that the people at the top of these lists are using assistance. Gathering PGN files over more than a decade, Professor Regan has been able to calibrate the expectations in these characteristics for players of a different rating.

Using this, Professor Regan can come up with a index, called Raw Outlier Index (ROI). 50 means you have met expectations, but anywhere between 40-60 is normal. 60-70 is still normal but take a complaint seriously. 70+ is a suggestion to contact the FIDE Fair Play Commission for further advice. In my years of over-the-board tournaments, I've only ever seen a ROI of 70 once, and that player was caught using assistance from an engine in Telford in 2018.

Sorted by this index, this is how the top 20 leaderboard with 4NCL's data over-the-board looks this season; we've taken out the names of the players!

Rank	Matc%	AvScD	ROI	#Mvs	Sc/#Gm
1	60.1%	0.117	62.6	163	3.5/ 4
2	64.5%	0.084	62.5	138	2.0/ 3
3	65.7%	0.040	61.2	105	3.0/ 4
4	64.0%	0.055	61.2	100	3.0/ 4
5	63.0%	0.051	60.9	165	3.0/ 4
6	61.9%	0.062	60.8	139	2.0/ 3
7	59.0%	0.086	60.8	134	3.0/ 4
8	55.3%	0.092	60.7	141	0.5/ 4
9	61.5%	0.124	60.6	179	3.0/ 5
10	63.6%	0.075	60.6	162	4.5/ 6
11	67.9%	0.042	60.6	109	3.0/ 4
12	51.9%	0.088	60.5	181	3.5/ 6
13	63.6%	0.058	60.2	165	4.5/ 6
14	58.0%	0.087	60.2	88	2.0/ 4
15	59.9%	0.077	60.2	187	3.5/ 6
16	55.8%	0.134	60.1	104	1.5/ 4
17	55.1%	0.098	60.0	127	3.0/ 4
18	58.7%	0.092	60.0	223	3.5/ 6
19	57.0%	0.094	59.9	165	3.5/ 4
20	63.0%	0.076	59.6	108	2.5/ 4

Professor Regan uses two engines, and some of these players are duplicates – the same player with both engines. There are 18 readings of 60 or higher, so should we be worried? No. The first comment is that it is clear that some players with an objectively poor scores are scoring highly; the system is score-independent. The most important comment is that there are 1097 names in this list. For a tournament of this size, we would naturally expect the screening to show a few players over 60. We would also expect to see that at the other end, there are players below 40; there are 26 such readings. We would expect to see the median ROI of 50; the median is actually 49.8, which means it is reasonable.

Here's the data from another tournament I was involved in:

Rank	Matc%	AvScD	ROI	#Mvs	Sc/#Gm
1	61.6%	0.101	67.0	250	5.0/ 7
2	58.9%	0.055	64.2	302	5.5/ 7
3	58.4%	0.082	64.0	250	5.0/ 7
4	71.9%	0.034	63.3	121	3.0/ 4
5	59.4%	0.136	62.2	106	2.5/ 3
6	47.7%	0.114	62.1	111	4.5/ 5
7	71.4%	0.031	62.0	119	2.0/ 3
8	68.5%	0.023	61.8	111	2.5/ 3
9	53.8%	0.075	61.4	78	2.5/ 3
10	54.0%	0.095	61.1	211	4.0/ 7
11	54.9%	0.130	61.1	266	3.5/ 9
12	52.9%	0.096	60.9	121	2.0/ 3
13	56.9%	0.090	60.9	216	5.0/ 7
14	52.4%	0.094	60.8	412	5.5/ 9
15	63.2%	0.049	60.7	190	6.5/ 9
16	54.8%	0.066	60.7	126	2.0/ 3
17	51.8%	0.114	60.6	394	5.5/ 9
18	50.4%	0.132	60.6	232	3.0/ 7
19	63.5%	0.049	60.4	203	6.0/ 8
20	56.1%	0.100	60.4	132	3.5/ 5
21	55.9%	0.069	60.4	202	4.0/ 5
22	52.3%	0.088	60.3	88	2.5/ 3
23	52.6%	0.095	60.3	291	4.5/ 7
24	62.0%	0.143	60.1	92	2.0/ 4
25	51.9%	0.119	60.1	131	1.5/ 3

Again, this is nothing to cause alarm. There were 1,364 readings in this one, and the median was 49.6, which is close enough to 50 to not worry about. There's a 67.0 at the top though – what is the explanation for that? That was in a Rapidplay tournament, and the problem was that the input was the player's FIDE Rapidplay rating. By making that adjustment, the ROI would drop accordingly and remove the cause for alarm.

Supplying any PGN file of any tournament tends to get results like this. I am sure that the editor would not approve of me supplying another 20 examples! You get the occasional glimpse of something like a ROI of 67, but having identified the reason for it, you can move on. So if we have this data for 4NCL Online, this is more or less what we would expect to see, with perhaps one case above ROI > 65 due to the size of the tournament. So what do we see?

Rank	Matc%	AvScD	ROI	#Mvs	Sc/#Gm
1	69.3%	0.030	74.0	189	7.5/ 8
2	76.5%	0.038	73.4	183	8.5/ 9
3	69.7%	0.049	73.3	175	7.5/ 8
4	70.6%	0.051	73.0	235	7.5/10
5	68.2%	0.064	71.6	220	7.5/10
6	70.1%	0.071	71.5	234	8.5/11
7	63.7%	0.069	71.5	267	8.5/10
8	62.7%	0.043	71.0	346	8.0/10
9	79.8%	0.033	70.4	119	2.5/ 3
10	73.1%	0.049	70.3	130	7.0/ 7
11	68.1%	0.053	70.3	251	8.5/11
12	78.0%	0.017	70.2	123	2.5/ 3
13	62.8%	0.041	69.7	188	6.5/ 7
14	73.5%	0.037	69.7	98	6.0/ 6
15	61.6%	0.057	69.6	185	7.5/10
16	65.4%	0.036	69.3	191	6.5/ 7
17	58.1%	0.050	69.2	265	7.5/ 9
18	60.1%	0.051	69.2	298	8.5/10
19	59.7%	0.052	69.1	320	8.0/10
20	71.7%	0.061	69.1	166	8.5/ 9

I've spoken to a number of cynics on the phone over the past three months, who don't believe that data can show the use of engine assistance. I would like to think that even the most hardened cynic might look at this and raise an eyebrow.

The median is 49.3, so the calibration is still OK. The MMP and ASD scores are noticeably different from over-the-board. But why do we suddenly have a string of people with a ROI over 70? Is there something about online chess?

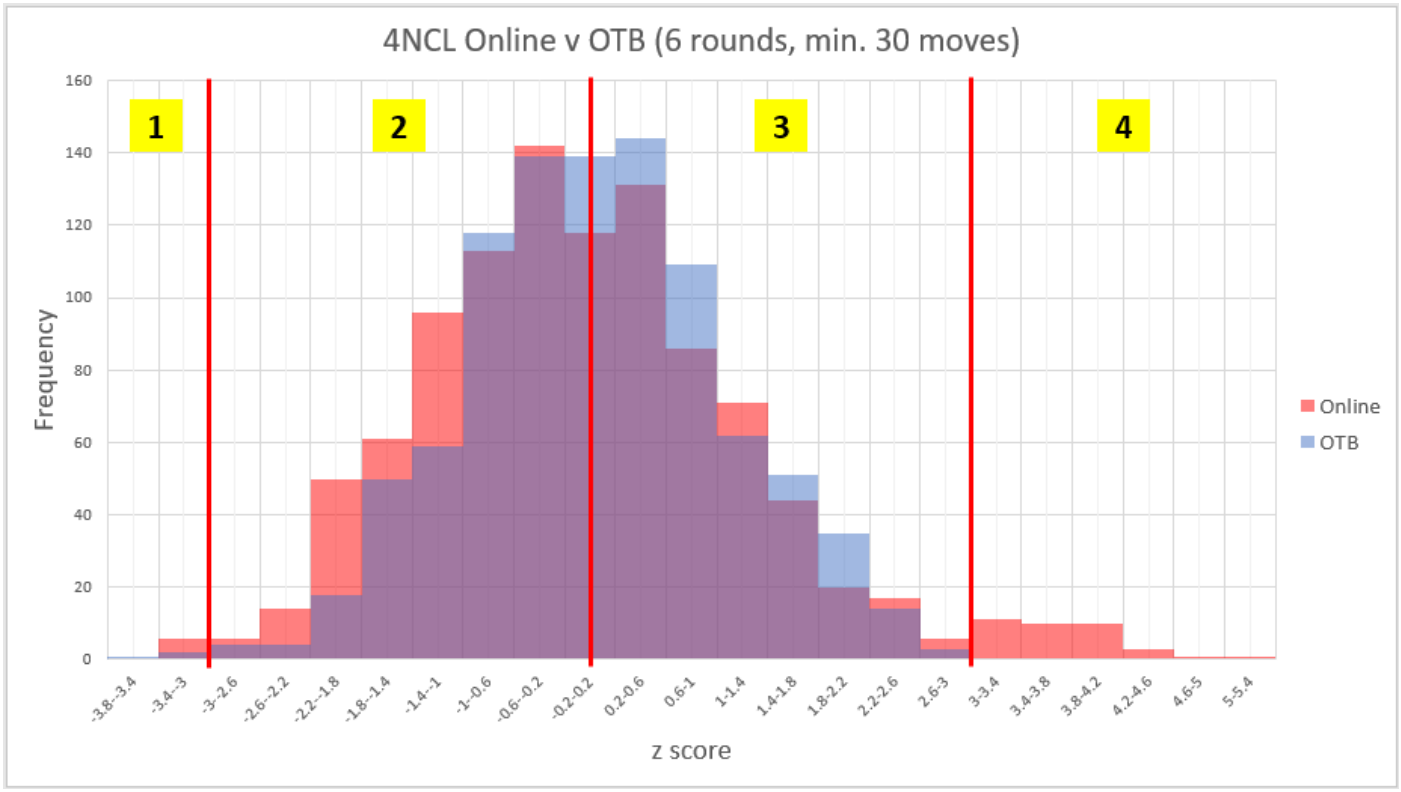
The ECF ran a Blitz tournament which lasted 24 hours, with several players playing more than 100 games. From a data-gathering exercise, it was a great format! Maybe if there is something special about online chess, we would expect to see something similar in the Blitz event to what we see in 4NCL Online.

Rank	Matc%	AvScD	ROI	#Mvs	Sc/#Gm
1	64.4%	0.040	84.8	205	24.0/29
2	44.0%	0.201	67.0	822	33.0/58
3	47.5%	0.151	63.6	255	10.5/16
4	47.5%	0.170	62.8	693	21.5/38
5	49.3%	0.159	62.8	795	32.0/42
6	48.3%	0.171	62.5	1272	37.0/57
7	48.1%	0.158	60.7	1400	70.0/84
8	42.5%	0.238	60.1	186	7.0/23
9	43.9%	0.166	60.0	1986	67.5/97
10	39.4%	0.203	58.4	738	25.5/51
11	45.8%	0.206	58.3	1212	36.0/71
12	90.0%	0.107	58.2	10	2.0/ 2
13	60.0%	0.076	58.1	15	1.0/ 1
14	47.9%	0.184	57.8	691	25.5/40
15	50.1%	0.140	57.5	387	14.5/19
16	42.4%	0.218	57.4	1866	50.0/92
17	39.8%	0.272	56.9	465	16.0/33
18	48.6%	0.130	56.6	541	23.0/26
19	58.8%	0.245	56.5	34	0.0/ 2
20	44.2%	0.175	56.5	852	18.5/32

What we see is far closer to what we see in over the board chess; there are fewer ROI > 60 cases because there are only a few hundred players involved. It's much harder to use engine assistance in blitz than classical chess due to the short time limit. Professor Regan's model is calibrated for blitz here, so the corresponding reduction in Move Match and ASD has been accounted for. The other thing to factor in is the fact that out of necessity, this event used the Chess.com ratings as an input, some of which were new accounts set up for the event, and several of these ROI > 60 can be explained by that. This seems to suggest that there is nothing special about online chess; you don't appear to suddenly play better, and it doesn't explain the numerous ROI > 70 we can see in 4NCL Online. The median of this data was 49.3, which is still fine.

A common thing I have heard over the past few months is that websites rush into flawed judgements with regard to issuing bans. My experience is the opposite. The PGN for this Blitz was sent off at 5pm on Sunday and the results came back later that Sunday evening. The player with a ROI of 84.8 still had their account open by the time I read the results email, and in fact they had used it to play other games on the website after the Blitz tournament had finished. Armed with this data, I manually reported the player, and within a few minutes the account was closed due to a fair play violation. In this specific case, I fully expect that the website would have done that itself once its checking system had caught up, but that intermediate step enabling the reporting to take place sped the process up. The burden of proof required by a website to automatically flag someone is high, and any anti-cheating regulations for an online tournament that rely solely on the provider without having Professor Regan's data to refine the search is going to have a lot of users who have used engine assistance slipping through undetected. I also reported the 67.0, which looks potentially suspicious, but the website took no action.

So far, I have only described Step 1 of the process. Step 2 of the process is Professor Regan's full test. This calculates a sigma score based on deviations from the expectation in the two characteristics described above. It also analyses the positions to greater depth, and only starts to check the positions when the opening book has been exhausted, rather than an arbitrary starting point of move 9. This is done on a case-by-case basis for up to four engines. However, with 4NCL Online this was happening much more often and it became impractical to do that. As a result, Professor Regan now supplies a spreadsheet with the z scores of all players in the competition, albeit using turn 14 as the start turn in order to avoid manually determining the end of the opening book for each game. After Round 6 of 4NCL Online, I received this spreadsheet for the first time, and plotted a graph of these scores against the scores for the 4NCL over-the-board. This was good timing, given the 4NCL 2019-20 season was paused after 6 rounds. The graph below shows the results, with any player who has played fewer than 30 moves in total ommitted.



I've added three red lines, at z=-3, z=0 and z=3. I've then divided the graph into zones 1-4. You can see the data is largely concentrated in zones 2 and 3, between -3 < z < 3. Both competitions have approximately 1000 players, so this is exactly what I would expect to see; you would expect a bit of "dribble" into zones 1 and 4. Zone 1 has a bit of dribble for both formats; although the worst zone 1 dribble in 4NCL over-the-board can be explained by a game-inputting mistake that Professor Regan discovered as a result of this work! Zone 4 makes the scale of the problem clear; the graph follows what is broadly a normal distribution, before suddenly rising up again and tailing off.

If a website flags or bans a player for using computer engine assistance in 4NCL Online, I inform the captain of the player's z score, and the MMP/ASD characteristics I described earlier. There have only been two confessions out of about 30 cases at the time of writing. The usual response I receive is that they have independently reviewed them and said that there isn't a shred of evidence to support the allegation. The resulting implication is that the website is wrong, and Professor Regan's model (which seems to have otherwise worked well for 10 years) is wrong, and then a generally dissatisfied-with-everything article is published somewhere.

I can't comment on how any platform's systems work, because I have not seen them and even if I were given access to it I would not be permitted to tell anyone. However, there are very few ways to skin this particular cat, and so I would expect that MMP and ASD are heavily involved in their detection processes. However, they also have other means of reaching the conclusion that are not contained in the PGN file.

This can take a z=3.5 player for the 4NCL, which looks dodgy but isn't enough to take action, and turn it into a z=5 or z=6 equivalent using their internal metrics.

So far I've discussed z scores without actually translating them into something more human to understand. It is a function of the normal distribution, and a z score can be translated into a probability. Professor Regan's model is an "honesty" test, which means that using external assistance is a deviation from that. z=4 means a 1 in 31,574 chance that no engine assistance was being used if this were the only data point under possible consideration. If we have 1,000 players in our tournament, then we would expect to see z=4 once every 31.574 tournaments; i.e. about once every 16 years given 4NCL Online is intending to run two tournaments per year. The highest z score we have had before being flagged at the time of writing is z=5.33, which means a 1 in 20 million that no engine assistance was being used. Using the same analogy, we would expect to see this naturally once every 10,198 years. Given this is an exercise purely in statistics, where do you draw the line?

In UK law, there are two legal tests: the "balance of probabilites" for a civil test, and "beyond a reasonable doubt" for a criminal test. One of the arguments put forward sometimes is that "You shouldn't do anything about cheating online unless you can be 100% certain." You can never be 100% certain, if a court applied this test then it would never convict anyone if the defendant didn't admit it. A line does need to be drawn somewhere, and so far as I am aware, chess websites actually work to "comfortable satisfaction", which is somewhere between "balance of probabilities" and "beyond a reasonable doubt". There are two organisations of note that also use this definition for similar cases – the Court of Arbitration for Sport, and the English Bridge Union.

A general rule of thumb I have learned during this competition is that websites tend not to take action if a player's z < 3.5. Professor Regan's z-scores are not part of the website's process, but their own tests will do the same sort of thing, plus factoring in the website-specific factors not contained within the PGN file. The higher the z score, the more likely you are to be flagged. In practice, if someone's z score is < 3.5 and they've been flagged by a website, investigation has revealed that it was based on games nothing to do with the 4NCL, or the 4NCL games were only a part of the overall package of evidence used.

An important thing to bear in mind is that a player's z-score does not correlate with their rating, for example, a Grandmaster won't get z=3 because they are a Grandmaster. If a Grandmaster plays like one, they'll get approximately z=0. On the other hand, if I play like a Grandmaster with my humble 1600 Elo rating, my z-score would be significantly > 0.

For this reason, it is important to get the input rating right for Professor Regan's system. In another capacity I am involved in, we know that English FIDE ratings are about right, so we would be within our rights to use them. However, we are aware of the human perception that English FIDE-ratings are stretched, but the statistics show that this is mostly because we are comparing them with countries in western Europe, where we can see that the ratings are comparatively inflated to ours. To avoid this, we use the highest possible rating we can find for someone, such as a national rating. The other cushioning is that the training data used by Professor Regan is over-the-board games of 2 hours each to move 60, whereas 4NCL Online games are 1 hours each to move 60. It is possible to measure the difference in the reduced ability between those two time limits due to the shorter time limit – in fact, this has been done for standardplay and rapidplay. However, we made the conscious decision to ignore the impact of the shorter time limit to build in a second layer of cushioning. This means that the z-score we give will be lower than it might otherwise have been. Statistically there is no reason for this cushioning, but it adds a human layer of protection to the tests.

The final point is that in my experience so far, humans are very poor judges of cases that correlate with the data. So far, every player who has been flagged by Lichess during 4NCL Online and hasn't confessed has been given a clean bill of health from either their team-mates or by their parents and coaches. There is only one player whose judgement on cases in 4NCL Online has correlated with the results of Professor Regan's system at all on a consistent basis, and he's in the world's top 100. If there is going to be a human element to this process, this is the level of player that needs to be involved in it. Sadly, even senior County Officials tend not to be of that level.

My advice for any online tournament looking to impose anti-cheating restrictions is:

- They should gather PGNs and send them off to Professor Regan
- They should use the screening results or z-score spreadsheet and report the problem cases to the website hosting the tournament
- Where the website unilaterally flags someone for using computer assistance, it is worth investigating if the data from Professor Regan supports the action they have taken; if not, the player can potentially use this to their benefit to try to get their flagging overturned
- Be prepared to receive a lot of grumpy, time-draining emails and phone calls on specific cases

If you run a tournament with a classical time limit and you are relying on the website unilaterally, then there will be lots of players with high z-scores who go undetected, and people using engine assistance will get away with it because the two-way process isn't properly being implemented. The result might be that fewer players were banned at the end of it, but that isn't really the metric of a successful online tournament. The success criterion is to catch the people who you are "comfortably satisfied" are using engine assistance, because unlike over-the-board, it is not really possible to prevent it.
