Gustav Mahlerplein 105-115 | 27th Floor | 1082 MS Amsterdam | The Netherlands

EUROPEAN PRINCIPAL TRADERS ASSOCIATION

FCA Occasional Paper "Quantifying the High-Frequency Trading 'Arms Race'" – Comments from FIA EPTA

The FIA European Principal Traders Association ("FIA EPTA"), the association that represents 28 trading firms in Europe that trade their own capital in futures, options, equities and bonds markets, is appreciative of the opportunity to provide an industry review and analysis of the **FCA Occasional Paper No. 50, "Quantifying the High-Frequency Trading 'Arms Race': A new methodology and estimates"** by Matteo Aquilina, Eric Budish and Peter O'Neill ("the Occasional Paper")[1]. The paper is highly relevant for our industry as it touches closely on areas our members are active in within the financial markets as market makers and liquidity providers.

Summary: Methodological Analysis of the Occasional Paper

FIA EPTA members welcome the FCA's commitment to commissioning data-driven analysis and believe that academic research plays a critical role in evaluating market structure, functioning, and integrity. In conducting this review of the Occasional Paper, FIA EPTA has conferred with its member experts and trading technology practitioners, who have shared their thoughts with FIA EPTA for collective feedback.

Notably, FIA EPTA members observe that the methodology applied by the authors of the Occasional Paper has several significant shortcomings. In line with this, FIA EPTA believes it is important to share our feedback regarding the Occasional Paper, including:

1. Shortcomings in the definition of "latency arbitrage"
2. Use of "stale" exchange message data not reflecting current market structure
3. Unsubstantiated extrapolation to global markets
4. Miscalculation of profit and loss (P&L)
5. Misleading time horizon
6. Errors in assumptions about the mechanics of how markets operate; and
7. Mislabeling the perceived impact as a "tax"

[1] Matteo Aquilina, Eric Budish and Peter O'Neill, Occasional Paper No. 50: "Quantifying the High-Frequency Trading 'Arms Race': A new methodology and estimates", at: https://www.fca.org.uk/publications/occasional-papers/occasional-paper-no-50-quantifying-high-frequency-trading-arms-race-new-methodology

1. **The Occasional Paper does not follow its own definition of "latency arbitrage"**.

The paper defines "latency arbitrage" as the "sniping" of a stale quote that a liquidity provider attempts to cancel. However, in the majority of purported "latency arbitrage races" that are measured in the paper, the liquidity provider never attempts to cancel its quote, indicating that it is comfortable with the quote being executed against.

If the paper solely measured instances where the liquidity provider actually attempted to cancel its quote, the number of "latency arbitrage races" would be reduced by approximately 70%.[2]

In addition, the paper found that these remaining "latency arbitrage races" are less profitable on average, corresponding to only about 20-30% of the total "latency arbitrage" profits measured in the paper.[3] This means that the size of the basis for the analysis is significantly overestimated.

2. **The Occasional Paper uses 'stale' stock exchange message data not reflecting current market structure.**

The paper's analysis is based on a 9-week period of LSE message data from 2015. However, the LSE has made a number of core system changes since that time including a change to ensure that public execution data is received at least as soon as private data. At the time of the sample, private data was received first giving the parties to each trade a key advantage over other market participants.
This advantage would have created "latency arbitrage races" that are no longer viable today, skewing the results.

This also means the paper is backward-looking rather than a useful insight into current market behaviour or outcomes.

3. **The Occasional Paper extrapolates findings from one exchange globally**.

The paper attempts to calculate profits from purported "latency arbitrage" on the LSE, and then extrapolates these numbers to the entire UK market and finally to global markets. The authors acknowledge that this methodology may have shortcomings; indeed, this is an unjustified extrapolation given:

- the wide variety of market structures and market participants around the world, and the fact that a significant amount of global trading activity occurs through trading protocols where "latency arbitrage" cannot occur (such as opening and closing auctions); and

[2] See Table 6.3 on page 47 (comparing (a) the column with 1+ cancel within the information horizon to (b) the baseline for purposes of the "races per day").
[3] Page 46.

- LSE only represents approximately 2.6% of the global value of share trading on electronic order books.[4]

It should be noted that trading strategies, including spread-crossing, depend on the structure of exchange order matching systems which differ across operators. In some matching systems, speed races such as the ones described in the paper are possible; in others they are not or are much less frequent. There are myriad differences among order matching systems and exchange network infrastructures. It is simply impossible to conclude that behavior observed on LSE must be present in the same way on other exchanges.

If the paper solely extrapolated to the UK market, the total profits from purported "latency arbitrage" would be GBP 60 million annually.[5] However, given closing auctions account for ~25% of daily volume, even this extrapolation is not viable.[6]

If the paper simply addressed these shortcomings, the total calculated profits from purported "latency arbitrage" would be dramatically reduced to approximately GBP 13.5 million annually across the entire UK market.[7]

4. The Occasional Paper miscalculates P&L and ignores trading costs.

In calculating the "profits" earned by "winners" of the races, the paper miscalculates P&L by measuring trade price to mid-market. This is seriously flawed because mid-prices are theoretical, non-tradable prices. A participant's P&L should instead be calculated based on the price at which the participant can actually close its position out in the live market.

If P&L were accurately calculated in this way, and transaction costs (not taken into consideration in the paper) would be accounted for, the outcome would far more often show participants losing money on trades than the paper is assuming.

5. The time horizon used in the Occasional Paper's analysis is too long.

The paper acknowledges that as the time horizon is lengthened, the core assumption that two market participants are "racing" based on the same information becomes strained, as the second participant may instead be reacting to additional information from the market.[8]

---

[4] 2018 data from https://www.world-exchanges.org/our-work/articles/wfe-annual-statistics-guide-volume-4.
[5] Page 55.
[6] Closing auctions accounted for 25% of daily volume in 2019: https://www.marketwatch.com/story/killing-time-the-closing-auction-is-eating-the-trading-day-2019-07-17
[7] See Table 6.3 on page 47 (comparing (a) the column with 1+ cancel within the information horizon to (b) the baseline for purposes of the "latency arbitrage tax, all volume (bps)" full sample, which shows a reduction from .42 bps to .13 bps. This percentage reduction is then applied to the GBP 60 million number referenced in footnote 4).
[8] Page 21.

The paper calculates that the distribution of "latency arbitrage races" has a mode of 46 microseconds.[9] Figure 4.1 in the paper shows the typical HFT reaction time between 29 and 50 microseconds (which includes exchange network latency). But the paper then uses as its baseline an "information horizon" that averages 202 microseconds and which may extend as long as 500 microseconds.[10]

By using too long a reaction time cut-off, the paper (a) significantly increases false positives, and (b) artificially increases the perceived profitability of races. By the authors' own admission, reducing this reaction time cut-off to a more reasonable 50 microseconds reduces the HFT "prize" by over 50%.[11]

Shortening the time horizon to 50 microseconds would reduce the total profits from purported "latency arbitrage" by a further 77% amounting to approximately GBP 3.1 million annually for the entire UK market.[12]

6. **The Occasional Paper makes no attempt to identify institutional or retail orders**.

The paper does not attempt to analyze the types of orders involved in "latency arbitrage races," including how many originate from institutional or retail investors, or from hedging activities that are generally recognized as beneficial.

The paper distinguishes two types of stylised participants: "investors" and "trading firms," and posits investors send orders "mechanically."[13] That is, they tend to be sent by (third party) execution algorithms. The assumption is that those orders are strictly aggressive. However, third party execution algorithms very often attempt to capture the spread and are therefore passive, not aggressive.

Moreover, professional trading firms provide liquidity and execute their liquidity demand in a variety of ways including third party execution algorithms and/or spread-crossing orders. Characterising liquidity provision as a passive-only activity is far from realistic. These orders are often indistinguishable from the "investor" type of activity.

As a result, the paper's implication that HFT firms are the sole beneficiaries of purported "latency arbitrage" may not be accurate as this may include other types of participants as well, including those transacting on behalf of end-investors.

[9] Page 25.
[10] Page 23.
[11] See Table 6.1.
[12] See Table 6.3 on page 47 (comparing (a) the column with 1+ cancel within 50 microseconds to (b) the baseline for purposes of the "latency arbitrage tax, all volume (bps)" full sample, which shows a reduction from .42 bps to .03 bps. This percentage reduction is then applied to the GBP 60 million number referenced in footnote 4).
[13] See page 15.

7. **A "tax" is not being assessed on end-investors as the Occasional Paper purports**.

The authors of the paper labels the $5 billion figure a "Latency Arbitrage Tax." This is very misleading. It is not evident that this competition is a zero-sum game, given the majority of purported "latency arbitrage races" measured in the paper include scenarios where the liquidity provider never attempted to cancel its quote, and therefore may have been satisfied with the execution.

Also, the paper recognizes that sometimes the sequence of events that makes up a race will happen by chance. One example of this is where a stock is trading on multiple venues and, in order to optimally access aggregate liquidity, an execution algorithm "sweeps" a number of venues i.e. simultaneously buy (or sell) on multiple venues. If participants react to an execution on venue A by placing orders on venue B, it may look like a race if they are received shortly after the execution algorithm's "sweep" order is received at venue B even if that order was in mid-flight at the time of the execution on venue A. As a considerable share of the volume in European equities happens in this manner, this distorts the results in a meaningful way.

It is normal that competition in markets rewards participants who make prices more informative. Liquidity provision, which can be provided through aggressive orders and immediacy of execution just as it can through passive orders and quoting, is a service to markets. The electronification of liquidity provision over the past decades has brought the average trading cost for investors down by more than 50 percent in the last 10 years. [14]

Further, the paper acknowledges that the winners and losers of the "latency arbitrage races" are often the same few firms. This means ordinary investors are not impacted by the niche "race" dynamic described.

Finally, as the measured effect of the races is much smaller than the minimum tick increment,[15] it calls into question the assumption that market-wide spreads would decrease if purported "latency arbitrage" was eliminated (rather than liquidity providers simply increasing their own profitability).

CONCLUSION

While we express the concern that the Occasional Paper suffers from serious methodological shortcomings, FIA EPTA and its members appreciate the intention of the study to analyse market structure functioning from the perspective of the end-user.

[14] Source: TABB Group Reg One Solutions, Effective/quoted spread for market orders of 100 – 1999 shares in the S&P 500. View the chart here: https://modernmarketsinitiative.org/wp-content/uploads/2016/03/Capture.jpg

[15] See Table 6.3 on page 47 (the implied impact in terms of ticks, using the column with 1+ cancel within the information horizon, is .05 ticks (.13/.99*.37)).

FIA EPTA members strongly believe that end-users should always be the main focus when assessing the effectiveness of markets. In that regard, FIA EPTA members consider that end-users of the markets, be they institutional buy side or retail investors, benefit when market structure facilitates efficient liquidity provision at low cost by ensuring that liquidity providers can compete on a level-playing field in transparent exchange-traded markets.

In taking the end-user perspective, we believe analysis should focus on the cost savings and operational efficiencies that end-users have enjoyed as a result of technological advances and automated trading. Numerous studies[16] indicate that the activities of electronic liquidity providers over the past 15 years have benefited European and global capital markets, making these fairer and more efficient for all participants. This has translated in lower cost to trade and better quality and immediacy of trade execution for end-users of the markets.

However, FIA EPTA members consider that further improvements to market structure for the benefit of end-users are still possible and strongly support further innovation to that end. FIA EPTA and its members are pleased, therefore, to be a further resource to the FCA, the authors of the paper, and any interested parties in relation to further market structure development.

[16] For an overview, see: The Economics of High-Frequency Trading: Taking Stock, at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2787542.