Comments to the Discussion Paper - Exploring safeguards in digital payments to curb frauds.

On 09 April 2026, the Reserve Bank of India released the **"Discussion Paper, Exploring safeguards in digital payments to curb frauds"** (hereafter **"Discussion Paper"**).

The Discussion Paper emphasizes the prevalence of social engineering fraud which continue to successfully cheat unsuspecting users out of money, despite there being system-level controls. The Discussion Paper focuses on making the digital payment landscape safer especially for those who may be more vulnerable to these ploys. It proposes four very specific interventions that seek to introduce positive friction in the payment value chain, allowing remitters some time to reconsider their payment decisions and hopefully breaking their 'hot state' of decision making.

In this response, we present our comments to the paper. We divide our comments into three sections.

- Section I: Offers a principle-level framework to reconceptualise fraud as a complex adaptive system (CAS) This reframing renders important lessons for the type of policy interventions that may be more effective than others in curbing fraud despite their resilient, ever-evolving nature.
- Section II: Unpacks each intervention proposed in the Discussion Paper. We discuss why each intervention may not work.
- Section III: Concludes with a 'model' mechanism that may bring enduring progress in dealing with fraud in the Unified Payments Interface (UPI) but also other aspects of the economy.

Section I: Reconceptualizing Fraud

We welcome the Reserve Bank of India's (RBI) focus and resolve on protecting citizens, especially those who may be more vulnerable from the rising prevalence of social engineered payment fraud especially in UPI. At the outset, social engineering attacks appear to be isolated, unconnected incidents that target individuals and do not pose system-level threats. However, looking closely at the typology of fraud broadly allows us to appreciate that characterizing fraud as a series of disconnected episodes may not be the correct formulation of the problem statement.

There is a growing consensus that fraud behave like or even may well be a CAS. CAS are characterised by the presence of multiple independent yet interconnected agents that constantly interact, learn, and adapt to each other, generating outcomes that cannot be predicted by studying the behaviour of a single actor alone.[1] Fraud as a system qualifies for this definition. Fraud is a consequence of several actors (such as the fraudster, the customer, the regulator, the financial service providers) constantly interacting, learning from and adapting each other's responses. New forms of fraud are likely to be more successful in their early days than when they have become well known because the defence architecture, such as the law enforcement bodies and the relevant agencies would have built guardrails and the customers would have revised their information set and become more cautious of a modus operandi gone viral. As the defenders and the customers come up to speed and become vigilant, the fraudsters evolve to exploit other unattended loops in the system, finding ever newer ways to inflict losses and make money.

This characterization of fraud has significant implications for policy designs. We enlist four, that we think are most relevant to the context at hand:

1. **Static variables such as demographic features of the remitter or transaction value have low-predictive power**. The adaptive nature of fraud and the low cost of improvisation indicates that fraud will almost always be successful in dodging interventions that are guided by static markers of risk. For instance, the moment a 70 year old citizen's account would be subject to greater scrutiny, fraudsters will gravitate towards cheating a 69 year old. Or they would invent newer ways to cheat the 70 year old, because that is the nature of the beast.

[1] https://www.sciencedirect.com/science/article/pii/S2949791425000740

Dynamic indicators learnt from the real-time forensic make-up of the transaction and enriched with the understanding of the wider context may be more useful triggers for policy interventions. For instance, marked changes in the velocity and behaviour of payment patterns can be a telling sign of fraud. Sudden, repetitive deductions that do not sit well alongside the regular transaction patterns can be more telling of the likelihood of fraud than the age of the remitter to whom it happens. A dynamic-threat based approach factors in established behavioural styles of the remitter (the time of the day, the device, the geography they usually transact), the relation between the remitter and the recipient (first-time or frequent, proceeded by a phone call) etc. and the wider context to determine the likelihood of a transaction being fraudulent. These indicators have the benefit of being able to detect and act on anomalies, a capability that static indicators sorely lack.
A review of literature on fraud prevention as an adaptive response yields a range of (illustrative) indicators that may have greater power of predicting at-risk individuals and suspicious transactions. We list these in Annexure-1.

2. **Adaptive policy interventions may fare better than the ones that are rigid and difficult-to-recalibrate**. Ideally, policy interventions should mimic the distribution of the fraud itself, intensifying both in quality and scale as target frauds pick up and recede from use or reduce in intensity when the fraud begins to retire. An in-built adaptiveness can reduce the burden of recalibrating through discrete interventions and control deadweight losses. Such adaptiveness, for instance, is seen in the stock market when circuit breakers are automatically triggered to intercept an extremely volatile episode. Similarly, dynamic credit commensurations that automatically expand when the relationship between the parties has crossed a reasonable trust threshold could be one way to imagine adaptiveness in policy interventions targeting fraud.

3. **A multi-stakeholder approach to securing each node in complex transactions.** As fraud become more sophisticated, they also usually become more layered. As seen in the landmark deepfake fraud in Hong Kong, a chief finance officer (CFO) of a multi-national corporation was persuaded to transfer USD 25 million by his senior management over a Zoom call, except everyone on the Zoom call except the CFO was a deepfake. This transaction points to lapses at several layers - the inability of a calling platform to gatekeep AI, the lack of internal safety valves that allowed a transaction of that size to be made through just a phone call, and an anomaly that would have shown up in the bank's activity as transactions of this size are unlikely unplanned or routine. Clearly, lapses at all ends led to a successful impersonation fraud. Therefore, it is imperative to adopt a multi stakeholder approach to instituting safety valves across the value chain that spreads far beyond the perimeter of the financial regulator.

4. **Harvesting information across sector and building AI on top of it to defend at scale.** As cybercrimes become cheap and scalable, they require live, adaptive, cheap and scalable counterweights. The growing research and practice of designing adaptive, dynamic fraud prevention tools using self-learning machine learning (ML) can rebalance the scales that heavily favor fraudsters today. The RBI's Digital Payment Intelligence Platform (DPIP) already embodies this principle, for example.

Section II: Input on the 4 interventions proposed in the paper

1. **Lagged credit for authorized push payments:** This option suggests that payments above ₹ 10,000 can be delayed by an hour, allowing the remitter to rethink the suspiciousness of the payment. However, this intervention may not have the desired effect:

- *First,* such safeguards are easy to circumvent. Fraudsters may restructure transactions below the limit of ₹10,000 thus only shifting the surface area[2] of the fraud rather than tackling its root cause.
- *Second,* safeguards such as introducing lag for transactions above ₹10,000 may introduce friction for legitimate transactions and risk casting an excessively wide net. In the context of UPI and instantaneous transactions, delay in credit may undermine the core value proposition of seamless and instant digital payments.
- *Third,* static safeguards are inherently limited in their ability to continuously learn and evolve alongside emerging fraud threats and risk obsolescence. Fraud patterns often evolve alongside technological adoption cycles. For instance, e-commerce related scams proliferated with the growth of online marketplaces, while more recently, digital arrest scams using AI-enabled deepfakes have emerged as significant fraud vectors. There is an inherent element of unpredictability in the manner in which fraud typologies evolve, and corresponding risks materialize. Thus, static safeguards by design are not equipped[3] to continuously learn from or respond to such evolving threats.
- International regulatory developments also reflect growing recognition of these limitations. In the United Kingdom, despite safeguards such as Confirmation of Payee and Reimbursement Scheme being in place, the fraud losses continue to remain substantial, with fraudsters adapting their techniques in response to existing controls. Consequently, the HM Treasury, the United Kingdom's Finance Ministry, has emphasized[4] the need for a more agile and outcomes-oriented framework to improve fraud prevention.

2. **Additional authentication for the elderly and Persons with Disabilities (PwD):** This option contemplates mandatory additional authentication through a trusted person for high value transactions undertaken by customers above the age of seventy and PwD. We appreciate that the proposed approach remains opt for all other individual customers, however, the way it is framed for specific demographics may warrant rethinking.

    - *First,* PwD itself is not a homogenous category. PwD have varying capabilities, and many are able to independently manage their financial affairs. The blanket assumption that PwD require support of a trusted person to authorize their payment transactions risks being paternalistic.
    - *Second,* making such authentication by trusted person mandatory rather than voluntary may be discriminatory and restrictive of their financial autonomy. Section 13 of the Rights of PwD Act, 2016 expressly recognizes the equal right of PwD to control their financial affairs[5].
    - *Third,* as mandatory imposition, this safeguard also encroaches upon the personal autonomy, privacy and dignity of PwD and those above the age of seventy. They will now have to prove their disability and age to enable additional authentication capability. Given the deeply personal nature of financial decisions, any safeguard proposed should preserve individual autonomy.
    - *Fourth,* vulnerability to fraud is not confined to demographic categories of age or disability. Social engineering frauds and authorized push payment frauds can exploit situational vulnerability, urgency, fear and information asymmetry, which can affect any customer, irrespective of the demographic.
    - *Fifth,* there may be situations where the trusted person is unavailable at the time of the transaction, causing delays in legitimate payments. Moreover, at the moment the approach

[2] https://www.eba.europa.eu/sites/default/files/2025-12/1709846a-84d9-47cf-86a0-b155efb34d66/EBA%20and%20ECB%20Report%20on%20Payment%20Fraud.pdf
[3] https://www.sciencedirect.com/science/article/pii/S1059056026000298
[4] https://assets.publishing.service.gov.uk/media/69ae77ddc78869bf8eb8a509/fraud-strategy-web.pdf
[5]https://www.indiacode.nic.in/bitstream/123456789/15939/1/the_rights_of_persons_with_disabilities_act%2C_2016.pdf

does not account for instances of coercion, misuse or collusion that involves the trusted person and makes them a vector of fraud themself.

3. **Ceiling the credit limit:** This option proposed introducing a regulatory measure to limiting aggregate credits in an account, beyond which additional credits to that account would require proof of a satisfactory business relationship between the customer and the bank. It contemplates this ceiling as ₹ 25 lakh, for annual aggregate credits into a bank account. The rationale is to control use of bank accounts as "mules" to route money proceeds of digital frauds. However, this approach may have certain limitations:

   - *First,* there may be legitimate transactions above proposed ₹25 lakh that are credited to a borrower's account in the ordinary course, such as disbursal of a home loan amount to an individual. In the absence of appropriate exceptions for such categories of genuine transactions, the proposed safeguard may operate disproportionately. Notably, the Discussion Paper itself contemplates exemptions for certain categories of large accounts, such as corporate and government accounts, in recognition of the nature and legitimacy of such transactions. A similar, calibrated approach would be required to operationalize this without which it risks creating unnecessary friction.
   - *Second*, the safeguard also does not cover relationships of customers with institutions other than banks, such as non-banking financial companies (NBFCs). NBFCs lend to financially excluded persons, low-income persons, or new to finance customers. These accounts of vulnerable persons have been known to be the target of fraudulent transactions, including being used as mule accounts.[6] By limiting the safeguard to customers who have relationships with banks, this effort for fraud prevention might overlook critical fraudulent transactions taking place outside of this ecosystem.

4. **Customer induced kill-switches:** This option proposes a kill switch capability across multiple payment channels such as UPI, cards, net banking, wallets and other digital instruments. The Discussion Paper suggests that certain transactions such as payment mandates and standing instructions could be exempted from customer-induced controls. The following points merit consideration:
   - *First*, this option shifts the burden back to the customer to prevent fraudulent transactions. The success of this option would be conditional upon factors such as digital and financial literacy of the customer and awareness of the customer about the fraudulent nature of the transaction.
   - *Second*, the option of a kill-switch may not be effective in cases where fraudster has the access to the device and the customer has been locked out.
   - *Third*, regulated entities differ considerably in terms of technological maturity, operational capacity and financial resources, which may influence their ability to implement this safeguard for fraud prevention.[7] While the larger regulated entities may possess the capacity to independently invest in fraud control infrastructure, smaller regulated entities may face greater operational and financial constraints in implementing comparable safeguards.

Using a multi-dimensional criterion, we tabulate the benefit and cost of each of these proposals:

Policy Intervention	**Benefit: Effectiveness in hindering the fraud**	**Costs: Social, economic or time costs**

[6] https://www.domain-b.com/companies-organisations/firms-companies/rbi-warns-nbfcs-of-money-mules-using-funds-for-money-laundering?utm_source=chatgpt.com
[7] https://arxiv.org/pdf/2506.12060

Lagged Payment	**Low** Smaller ticket-size fraud in batches may become common	**High time cost for the customer** **Greater false positives**
Trusted individuals for authentication	**Low** Fraudsters may disproportionately target customers just below the 70-yr old threshold	**Potentially discriminatory** **Raises issues of liability,** when the trusted individual approves a fraudulent transaction
Credit commensuration	**Low** Fraudsters will be able to by-pass this by increasing the number of mule accounts they use to launder and move money.	**Setback to the ease of doing business** Unclear how big-ticket loans/ insurance sums/ real estate transactions will be treated
Customer-led kill switch	**Moderate** But it relies on the customer to beware	**Unclear how device takeover (and similar) frauds will be dealt with**

Section III: Recommendations and Conclusion

We humbly submit that the effectiveness of these proposals and the general stance to fraud prevention may benefit from the following:

1. **Anchoring policy interventions in transaction-based data**. By predicating interventions on the merit of the transaction instead of the demographic features, the effectiveness of options 1 and 2 can be greatly enhanced. For instance, instead of triggering the lag for transactions above ₹10,000, if the lag got triggered in response to unusual transaction pattern such as repeated small-value deductions, unusual behavior patterns such as the transaction being clearly initiated from a different geography than the remitter's usual location, etc., the likelihood that the intervention is responding to a fraud incident would rise substantially. Similarly, instead of every transaction made by a particular group, if all transactions that are assessed suspicious based on real-time, transaction and contextual data with proven predictive powers, could be subject to additional authentication, that may again increase the effectiveness of the intervention. To summarize, shifting focus from static risk indicators to dynamic risk indicators can lead to sharper targeting and interception of fraudulent transactions.

2. **Making these policy interventions universal and voluntary (where applicable).** Policy interventions such as those proposed by options 1 & 2 are amenable to this principle. If those options can be offered to all users, on a voluntary basis, irrespective of their demographic features such as age and disability, that may also enhance the efficiency of the system. Static rules for one may not be effective in reducing the surface area of the attack, they may only incentivise fraudsters to look for newer methods and demographics. Second, they have a tendency to flag and restrict the vast swathe of legitimate transactions causing deadweight losses and frustration in the system. Therefore, moving away from using static indicators to target interventions, making them voluntary, would allow vulnerable populations to self-select relevant safeguards.

3. **The capability for dynamic, adaptive targeting, interception already exists.** RBI initiatives such as the DPIP and MuleHunter are designed to harvest and act upon dynamic indicators of vulnerability. These initiatives seek to use AI & ML and continue to monitor and learn from evolving modus operandi which increases their predictive power. These infrastructure exhibit several properties necessary for responding to CAS like fraud. These include real-time risk scoring of the transaction based on behavioural patterns (rather than static indicators); centralized fraud intelligence sharing across banks,

payment systems, and NBFCs; negative registries of mule accounts; fraudulent phone numbers, repeat offenders and pre-transaction alerts to help banks block suspicious payments before completion.

The most severe constraints that regulators face in implementing adaptive regulation include an appreciation for the need of adaptiveness, developing the capability of creating information gathering mechanisms and training AI on them[8]. However, by designing the DPIP and the MuleHunter the regulator has already exhibited the capability to imagine and design for adaptive regulation. Therefore, by making these efforts more central to the fraud detection mechanism, using the risk score awarded by the DPIP and not static indicators of age or transaction value could yield unmatched dividend and lead to a resilient payment ecosystem.

---

[8] https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/adaptive-regulation/10984D9EE8FE563187E0F28AB31FA6EA

Annexure-1

Dynamic indicators of fraud based on the literature on feature engineering

1. Temporal & Behavioural Velocity Markers

- Transaction Frequency Spikes: Tracking the number of transactions within a short timeframe (e.g., last 1 hour, 24 hours).
- Rapid Successive Transactions: Detecting high-speed, repetitive small-value transfers, common in testing stolen credentials.
- Velocity Changes: Sudden deviations from a user's historical average transaction count or amount. [1, 2, 3, 4, 5]

2. Behavioural Biometrics & User Context

- Keystroke/Gesture Dynamics: Analysing touch patterns, speed of PIN entry, and navigation patterns on the app to detect non-human or coerced behavior.
- Time-of-Day Anomalies: Transactions occurring at unusual times (e.g., 3:00 AM) that do not fit the user's profile.
- Amount Distribution Analysis: Identifying amounts that do not match the user's typical spending patterns or are just below the mandatory PIN verification threshold. [1, 2, 3, 4, 5]

3. Device & Network Fingerprinting

- New Device Identification: Flagging transactions from new device IDs, IP addresses, or sudden changes in device fingerprint.
- Geolocation Discrepancies: Identifying when transaction locations do not match the user's usual area or are far apart within a short time ("impossible travel").
- Remote Access Detection: Identifying screen-sharing apps or remote access tools (RATs) running in the background, a high indicator of "mule" account manipulation. [1, 2, 3, 4, 5]

4. Network-Based & Social Graphs

- Payee/Payer Risk Profiling: Calculating the "risk score" of the recipient or sender based on their history with other flagged accounts.
- First-Time Interaction Risks: Identifying and flagging P2P transactions between a buyer and seller who have never transacted before.
- Mule Account Indicators: Identifying accounts that receive funds and instantly transfer them out (money laundering). [1, 2, 3, 4, 5]