INSTITUTIONAL DETERMINANTS OF PRIVACY DESIGN IN CENTRAL BANK DIGITAL CURRENCIES

Martin Bok – Dominik Stroukal

Abstract

Central bank digital currencies (CBDCs) make transaction privacy an institutional design choice. We ask whether political and economic conditions predict the privacy posture documented by central banks, distinguishing privacy vocabulary from design commitments. A frozen multilingual corpus contains 113 official or authority-attributed documents (3,963 pages) across forty-seven jurisdictions and two institutional composites. Two independently run large-language-model extractors produced 11,358 source-verified statements, deduplicated into 6,949 candidates and aggregated by a predeclared equal-vote ensemble. On a separate calibration set, the verified union attained precision 0.980, recall 0.766 and span fidelity 1.000; because recall missed the predeclared criterion, the analysis remains exploratory for now. In a blind probability sample, reviewer-specific design-weighted validity was 0.786 and 0.761. Inclusion agreement was 62.6% (Gwet's AC1 = 0.426; Krippendorff's $\alpha = -0.069$), and exact-code agreement among 214 jointly retained candidates was 46.7%. In the production union, privacy-family share and documented privacy posture correlate at $r = 0.219$. No tested institutional condition survives family-wise correction; the largest posture-side association is with the informal economy ($r = -0.286$). The design demonstrates both the value and current limits of reproducible multi-model LLM qualitative extraction.

Key words: central bank digital currency; privacy; institutional design; csQCA; financial infrastructure

JEL Code: E42, E58, C83

Introduction

Existing trackers record project stage and some technical attributes, while central-bank surveys aggregate self-reported motivations (Kosse & Mattei, 2022; Auer et al., 2023); neither systematically codes project-level design commitments or their privacy implications. CBDCs allow public authorities to reconfigure consequential properties of FIAT money, including programmability and access to transaction data, so projects at the same stage may embody very different institutional choices. This paper therefore measures documented design rather than project status. Privacy is particularly consequential. Physical cash normally settles without

creating a central record of who paid whom. A retail CBDC may reproduce part of that boundary or make every transaction visible to intermediaries and public authorities. The outcome is therefore a policy choice, not a technical inevitability.

Payment privacy has welfare content. Cash protects buyers from routine identification (Kahn et al., 2005), and privacy can have public-good properties that private payment markets underprovide (Garratt & van Oordt, 2021). CBDC design nevertheless combines this objective with financial integrity, monetary control and operational resilience. Agur et al. (2022) formalise the trade-off between cash-like and deposit-like features; Auer and Böhme (2020) show how architecture and access shape data exposure. Cross-country differences should therefore be observable in official design documents.

Hypotheses

The political-economy literature implies four directional expectations for the institutional conditions used below. Democratic accountability should raise the political cost of state access to transaction data, while central-bank independence may insulate long-run monetary and institutional credibility from short-run surveillance demands. Conversely, a larger informal economy and higher money-laundering risk may strengthen enforcement incentives for traceability (Auer et al., 2023; Cukierman, 2020). The earlier crisp-set qualitative comparative analysis (csQCA) framing treats such conditions conjuncturally; the present MSED analysis tests the four marginal implications visible in Figure 2:

H1. Liberal democracy is positively associated with a more privacy-protective documented CBDC posture.

H2. Central-bank independence is positively associated with a more privacy-protective documented CBDC posture.

H3. A larger informal economy is negatively associated with documented CBDC privacy posture.

H4. Higher Basel AML risk is negatively associated with documented CBDC privacy posture.

Because privacy vocabulary can diverge from architecture, each hypothesis is also evaluated against the privacy-family vocabulary measure as a diagnostic. Table 2 reports ensemble estimates with bootstrap intervals and Holm-adjusted p-values; Figure 2 shows provider sensitivity. These are associational tests, not causal or necessity/sufficiency claims.

The empirical difficulty is measurement. A document may promise privacy while preserving full state visibility; counting privacy words then measures framing, not informational exposure. Text may also discuss alternatives, cite external research or report a pilot without adopting a feature. Large multilingual corpora intensify these problems: manual census coding is costly,

while unconstrained model output can miss eligible passages, invent quotations or confuse proposals with decisions. Multilingual models can apply one codebook directly to source-language text, but performance remains model- and task-dependent (Rathje et al., 2024). Reliability therefore depends on explicit codebooks, validation and audit trails (Gilardi et al., 2023).

We address the substantive and methodological questions together. We build separate measures of privacy-family prevalence and documented privacy posture, process the same frozen sources independently with two providers, and preserve disagreements through an equal-vote ensemble and provider-specific sensitivity analyses. A probability-sampled blind human audit estimates candidate validity and coding agreement. The institutional analysis is descriptive rather than causal.

1 Data and methods

1.1 Corpus and unitisation

The frozen corpus contains 113 official or authority-attributed CBDC documents, 3,963 pages in total, covering forty-seven national or monetary jurisdictions and two institutional composites. It includes 101 English-language documents and 12 documents (236 pages) in Spanish, French, Japanese and Russian/Ukrainian. Sources include design reports, consultation papers, technical documents, legislation and official project updates. Each PDF, page image and extracted text unit is identified by a SHA-256 hash; one reference-only document is excluded from extraction.

Pages are the base unit and are split only when they exceed the frozen character budget; image renders accompany unreliable text layers. All 3,963 units are enumerated, and no semantic prefilter is applied. The codebook contains 35 labels covering privacy, AML/KYC design, architecture, programmability, remuneration, interoperability, access, resilience, legal basis and project learning. Accepted statements require an exact quotation, primary code, observation/decision/recommendation status (ODR), applicable privacy direction and relation, and strength from 1 to 3.

1.2 Two-model extraction and source verification

OpenAI gpt-5.6-terra and Anthropic claude-sonnet-5 received byte-identical source packages, prompts, codebooks and output schemas. They operated independently and were blinded to each other's output. JSON outputs were schema-validated, quotations were normalised and matched to the source, and candidates were deduplicated only when authority, source unit and

normalised span matched. Provider identity and labels remain archived so that sensitivity analysis is reproducible.

A model-produced statement entered the canonical set only if it was a complete claim about the named authority's CBDC project: an adopted or rejected decision, explicit issuer proposal or requirement, concrete feature, executed-pilot finding, or supported project-specific risk, trade-off, limit or operating rule. Headings, definitions, stakeholder lists, consultation questions, generic context, cited foreign research, future-research statements, unadopted hypotheticals and fragments were excluded. The smallest complete quotation had to be recoverable from the frozen source; one source-only retry was permitted, and unsupported spans were rejected before matching and deduplication.

Before production, the exact protocol was tested on 78 calibration pages. OpenAI achieved precision 1.000 and recall 0.625; Claude achieved 0.977 and 0.656; the source-verified union achieved 0.980 and 0.766. Span fidelity was 1.000. The union missed 15 of 64 probability-sample positive paragraphs, and its document-cluster recall interval was 0.633–0.881. These results passed the precision and span gates but did not meet the frozen recall thresholds. The sealed 40-page reserve was therefore not opened. Production output is analysed as an exploratory, source-verified candidate dataset with sampled human validation, not as a validated automated census.

Table 1: Extraction, agreement and validation status

Quantity	OpenAI	Claude	Verified union / design
Production output	5,675 statements / 5,622 candidates	5,683 / 5,634	11,358 statements; 6,949 unique candidates
Calibration precision	1.000	0.977	0.980
Calibration recall (95% CI for union)	0.625	0.656	0.766 [0.633, 0.881]
Calibration span fidelity	1.000	1.000	1.000
Exact code-set agreement	n/a	n/a	41.5% among 4,307 shared candidates
Human probability sample	n/a	n/a	365/6,949 (5.25%), both authors
Human candidate validity (95% CI)	78.6% [75.5, 81.6]	76.1% [73.8, 78.3]	77.3% [75.3, 79.3] descriptive midpoint
Human agreement / dual-empty audit	Exact code 46.7%; family 53.3%	ODR 59.3%; privacy fields 90.2-91.1%	Inclusion 62.6% (alpha -0.069; AC1 0.426); 1/31 missed unit, 2 claims

Source: authors' computation.

1.3 Ensemble rule and sampled human validation

Every deduplicated candidate contributes total analytic mass 1. A sole provider receives that mass. Where both providers emit the candidate, each receives 0.5; a provider's share is divided equally among its distinct codes, and repeated evidence for one code cannot increase weight. Conflicting categorical labels are retained as fractional votes, never averaged into a synthetic

category. Decisions receive ODR weight 1 and proposals 0.5; findings remain in descriptive distributions but not in commitment scores. The ensemble is the primary analysis. Complete OpenAI-only, Claude-only and exact-code-consensus variants are robustness checks.

Human validation uses a deterministic, stratified probability sample selected before coding. Both authors independently reviewed the same 365 candidates (5.2526% of the union), stratified by language and provider origin, plus 36 of 359 units where both models returned no claim. Provider identity and machine labels were hidden, review files were locked before comparison, and blind status was confirmed. One candidate without a usable translation was excluded, leaving 364 analyzable cases. Population estimates use inverse inclusion probabilities; confidence is treated as high and exclusion reasons are not analysed. The dual-empty audit was conducted jointly after candidate coding; five untranslated units were excluded. The descriptive midpoint is reportable before adjudication but is not a final consensus estimate (Krippendorff, 2019).

1.4 Measures and statistical specification

The vocabulary measure is the weighted PRIV-family share among privacy, AML, KYC and technical families. Documented privacy posture is a signed aggregate of source-anchored design choices: direction is positive, negative, conditional or neutral; visibility from the state receives full relation weight and visibility from intermediaries half weight; ODR, strength and ensemble allocation multiply the contribution. Six purpose centres (cash substitution, financial inclusion, payment modernisation, monetary transmission, sovereignty and competition, and state control) combine the frozen code map with lexical signals. Ensemble percentile thresholds calibrate all provider variants on the same scale.

We report pairwise-complete Pearson correlations with the V-Dem v15 liberal-democracy index for 2024 (Coppedge et al., 2025), central-bank independence (Romelli, 2025), informal-economy size (Elgin et al., 2021), and the Basel AML risk score (Basel Institute on Governance, 2024). Confidence intervals use 10,000 bootstrap draws, p-values use 10,000 permutations, and Holm correction covers the eight primary institution–measure tests. Associations are descriptive; the small and selected jurisdiction sample does not identify causal effects.

2 Results

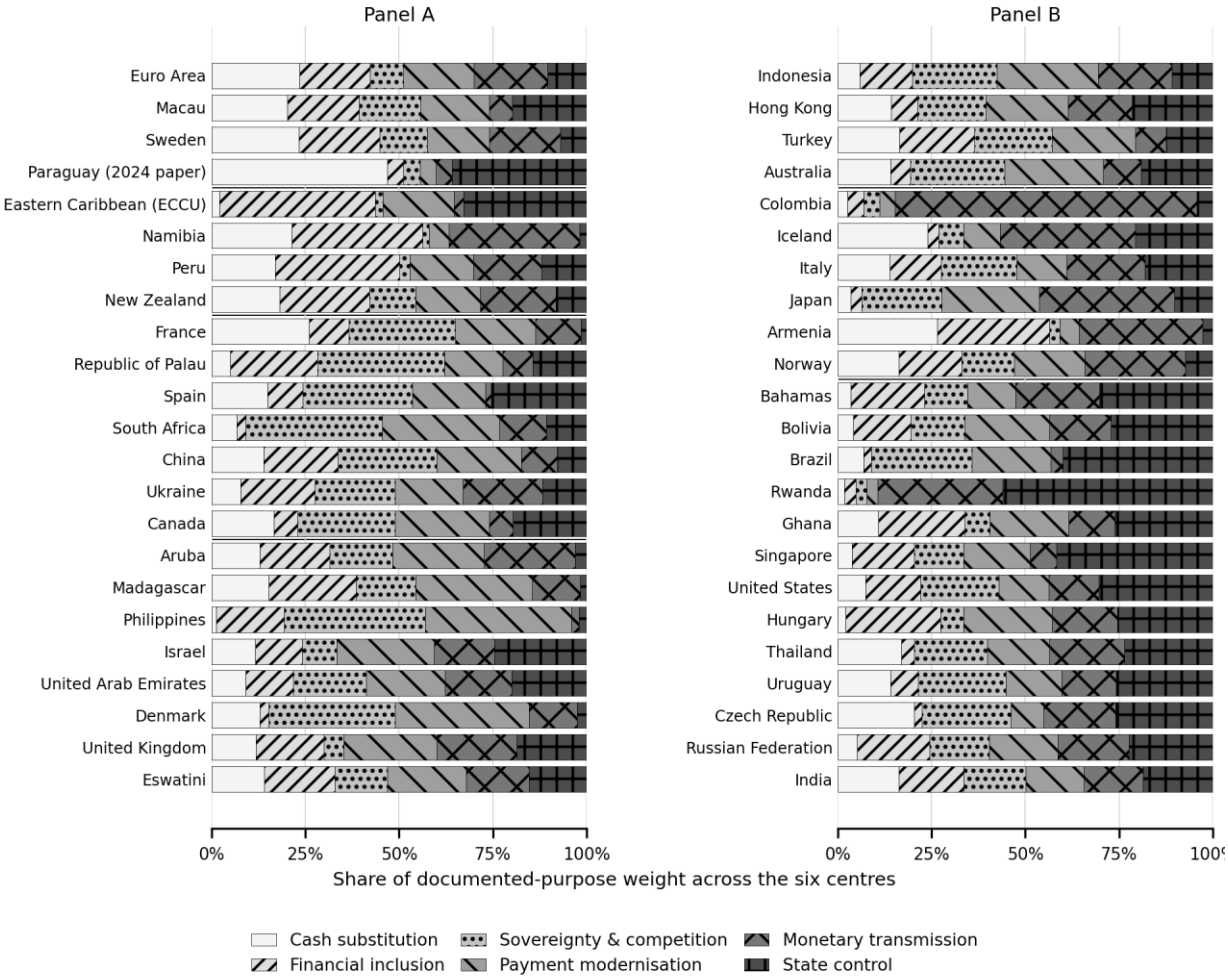
2.1 Model overlap and composition

The providers agree on the exact code set for 1,788 of 4,307 shared candidates (41.5%), on the code family for 44.9%, and on ODR status for 62.8%. Single-provider candidates represent 38.0% of the 6,949-candidate union. These are not error rates: a candidate may be valid despite

being found by only one model, and two codes may both be defensible. They do show why consensus-only analysis would discard substantial evidence and why silently selecting one provider would embed provider-specific structure.

Figure 1 presents the ensemble by jurisdiction. State control is the dominant centre for 14 entities and payment modernisation for 12, but these aggregate counts conceal substantial within-project mixtures: most projects distribute non-trivial weight across several centres, and many have a close second-ranked purpose. The dominant centre is therefore a summary of continuous composition rather than an exclusive project type.

Figure 1: Country-level composition of documented CBDC purpose



Source: authors' computation. Each horizontal bar normalises the six v10.2e ensemble centre scores to unit mass. Panels A and B contain all 46 country jurisdictions; the institutional composite is excluded.

2.2 Privacy posture and institutional conditions

Privacy vocabulary and documented posture are weakly related. In the ensemble their correlation is r = 0.219 (n = 39; 95% bootstrap CI [−0.102, 0.525]), so privacy-family frequency explains only 4.8% of posture variance. Provider variants remain modest (r = 0.195 OpenAI; 0.315 Claude). The distinction is therefore empirically consequential: a document's attention

to privacy is not a reliable proxy for the informational boundary its proposals and decisions create.

Viewed against H1–H4, the evidence is directional rather than confirmatory (Table 2). H1 receives weak point-estimate support: liberal democracy is positively related to vocabulary (r = 0.210) and posture (r = 0.152), but both intervals include zero. H2 is not supported because central-bank independence is slightly negative on both measures (r = −0.033 and −0.099). H3 is the most consistent pattern: informal-economy size is negative for vocabulary (r = −0.216) and posture (r = −0.286), but the posture interval touches zero and its Holm-adjusted p-value is 0.651. H4 appears only in language: Basel AML risk is negatively related to vocabulary (r = −0.292) but essentially unrelated to posture (r = 0.002). No hypothesis survives family-wise correction.

Table 2: Ensemble correlations with privacy vocabulary and posture

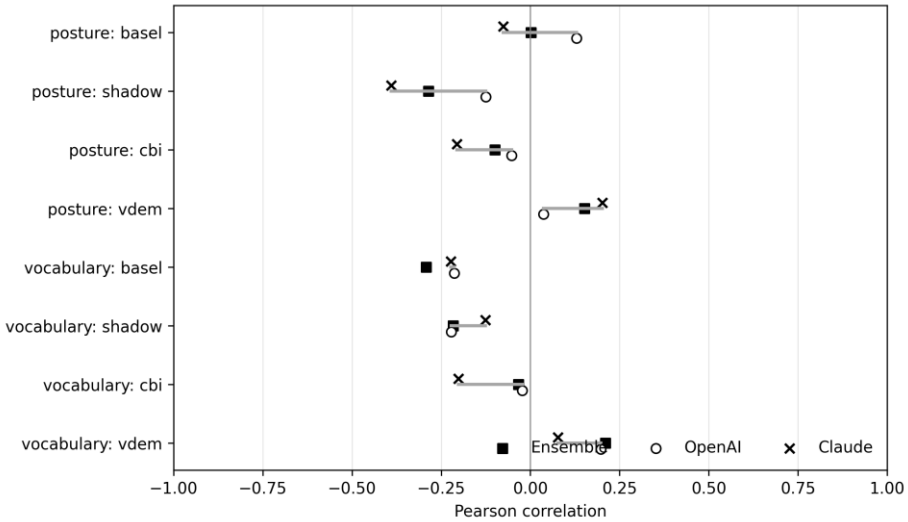
Measure	Condition	r	95% bootstrap CI	Holm p
Vocabulary	Liberal democracy	0.210	[−0.137, 0.488]	1.000
Vocabulary	Central-bank independence	−0.033	[−0.392, 0.336]	1.000
Vocabulary	Informal economy	−0.216	[−0.515, 0.150]	1.000
Vocabulary	Basel AML risk	−0.292	[−0.582, 0.016]	0.748
Posture	Liberal democracy	0.152	[−0.157, 0.417]	1.000
Posture	Central-bank independence	−0.099	[−0.561, 0.297]	1.000
Posture	Informal economy	−0.286	[−0.532, 0.007]	0.651
Posture	Basel AML risk	0.002	[−0.235, 0.255]	1.000

Source: authors' computation; pairwise n = 31–38. Permutation p-values are Holm-corrected over eight primary tests.

2.3 Provider sensitivity

Figure 2 makes the provider sensitivity of H1–H4 visible. Most institutional correlations remain small under both providers, but some ranges are substantively wide. The H3 posture–informal-economy correlation is −0.125 under OpenAI and −0.390 under Claude (ensemble −0.286). The H1 vocabulary–democracy estimate ranges from 0.077 to 0.198, while the selected consensus-only subset reverses sign (−0.203). The ensemble therefore remains primary, while provider and consensus variants bound model dependence.

Figure 2: Provider sensitivity of institutional correlations



Source: authors' computation. Squares are ensemble estimates; circles and crosses are provider-only estimates. Grey lines show the provider range. Rows correspond to H1 (V-Dem), H2 (CBI), H3 (shadow economy) and H4 (Basel AML risk), each shown for vocabulary and posture.

2.4 Blind human validation

The two reviewers produced similar aggregate estimates but limited item-level reproducibility. Design-weighted candidate validity was 78.6% for the main author (95% CI [75.5%, 81.6%]) and 76.1% for the second author (95% CI [73.8%, 78.3%]); their pre-adjudication midpoint was 77.3% (95% CI [75.3%, 79.3%]). They agreed on inclusion for 228 of 364 analyzable candidates (62.6%; IPW 62.6%), with Gwet's AC1 = 0.426 and Krippendorff's nominal $\alpha = -0.069$. The coefficients diverge because both reviewers retained most candidates (79.1% and 75.8%), creating a prevalence, or kappa-type, paradox: alpha's chance-agreement correction exceeds the observed agreement, whereas AC1 uses a less prevalence-sensitive correction (Gwet, 2008; Krippendorff, 2019). Reporting both therefore reflects different chance corrections rather than a computational error. Among 214 jointly retained candidates, exact-code agreement was 46.7%, family agreement 53.3%, and ODR agreement 59.3%; privacy direction and relation were 91.1% and 90.2%. The joint dual-empty audit found one missed unit among 31 analyzable units (3.2%; Wilson 95% CI [0.6%, 16.2%]) and recovered two eligible claims. The results support an approximately three-quarter candidate-validity rate while showing that the inclusion boundary and fine-grained taxonomy require adjudication and refinement.

3 Discussion

The current evidence does not support a strong regime account of CBDC privacy. Democracy estimates are small and uncertain, while informal-economy size is the largest structural

correlate but does not survive correction. Institutional narratives inferred from a few visible projects are therefore not recovered robustly in this larger corpus.

Methodologically, multilingual LLM extraction is measurement rather than clerical automation. Source verification suppresses unsupported evidence but cannot guarantee that every eligible passage was found or correctly classified. Calibration recall was 0.766, below the predeclared criterion, and exact-code agreement was 41.5% among shared model candidates. The blind human audit sharpens this boundary: the reviewers' aggregate validity estimates were close, but raw inclusion agreement was 62.6% and exact-code agreement among jointly retained candidates was 46.7%. The equal-vote union therefore preserves coverage while provider-specific and consensus variants expose model dependence; human adjudication remains necessary for a single consensus estimate.

Five limitations remain. Corpus coverage depends on document availability and central-bank transparency, underrepresenting early-stage and low-capacity jurisdictions. Official publications may not predict implementation. The probability-sampled audit estimates candidate validity and agreement but does not relabel the full machine-generated union. Its 77.3% midpoint is pre-adjudication and must not be interpreted as a final consensus estimate. The dual-empty audit covers only units where neither model found a claim and was reviewed jointly by the authors rather than as a second independent inter-rater sample. A complete production-recall estimate and strong causal claims are therefore missing.

Conclusion

CBDC privacy cannot be inferred reliably from how often official documents use privacy language. In the v10.2e ensemble, vocabulary and documented posture share less than 5% of variance, and none of four political or economic conditions survives family-wise correction. The informal economy produces the largest posture-side association, but it remains a hypothesis for larger and better-balanced samples. The result is not that institutions never matter; it is that the present public record does not identify a stable macro determinant.

The larger contribution is a reproducible boundary around LLM-assisted qualitative research. Two independent providers, frozen source packages, exact-span verification, archived raw outputs, an explicit ensemble rule and a probability-sampled blind human audit turn model output into inspectable evidence. The calibration and human results also define the method's boundary: high precision did not compensate for recall below the predeclared criterion, and similar aggregate validity estimates did not imply strong item-level agreement. The substantive tables retain the full union and disclose provider sensitivity rather than erase disagreement. This

project remains work in progress. Future research will test whether stronger LLMs, a larger and better-balanced document sample, and a more precise family structure improve recall, cross-provider stability and human coding agreement.

Generative AI disclosure: OpenAI GPT-5.6 and Anthropic Claude Sonnet 5 assisted extraction, content generation and proofreading. Source spans were automatically verified and sampled outputs human-audited. The authors take responsibility for the manuscript.

Data and code availability

Frozen code, manifests, canonical provider outputs and results are available at https://github.com/BokMartin/cbdcdataset; raw API archives are not redistributed, but their hashes and run metadata are retained. Project website: https://cbdcdataset.org/.

References

Agur, I., Ari, A., & Dell'Ariccia, G. (2022). Designing central bank digital currencies. *Journal of Monetary Economics, 125*, 62–79. https://doi.org/10.1016/j.jmoneco.2021.05.002

Auer, R., & Böhme, R. (2020). The technology of retail central bank digital currency. *BIS Quarterly Review*, March, 85–100.

Auer, R., Cornelli, G., & Frost, J. (2023). Rise of the central bank digital currencies. *International Journal of Central Banking, 19*(4), 185–214.

Basel Institute on Governance. (2024). *Basel AML Index 2024: 13th public edition: Ranking money laundering and terrorist financing risks around the world*.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Angiolillo, F., Bernhard, M., Cornell, A., Fish, M. S., Fox, L., Gastaldi, L., Gjerløw, H., Glynn, A., Good God, A., Grahn, S., Hicken, A., Kinzelbach, K., Krusell, J., ... Ziblatt, D. (2025). *V-Dem Country-Year Dataset v15*. Varieties of Democracy Project. https://doi.org/10.23696/vdemds25

Cukierman, A. (2020). Reflections on welfare and political economy aspects of a central bank digital currency. *The Manchester School, 88*(S1), 114–125. https://doi.org/10.1111/manc.12333

Elgin, C., Kose, M. A., Ohnsorge, F., & Yu, S. (2021). *Understanding informality* (CEPR Discussion Paper No. 16497). Centre for Economic Policy Research.

Garratt, R. J., & van Oordt, M. R. C. (2021). Privacy as a public good: A case for electronic cash. *Journal of Political Economy, 129*(7), 2157–2180. https://doi.org/10.1086/714133

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29–48. https://doi.org/10.1348/000711006X126600

Kahn, C. M., McAndrews, J., & Roberds, W. (2005). Money is privacy. *International Economic Review, 46*(2), 377–399. https://doi.org/10.1111/j.1468-2354.2005.00323.x

Kosse, A., & Mattei, I. (2022, May). *Gaining momentum: Results of the 2021 BIS survey on central bank digital currencies* (BIS Papers No. 125). Bank for International Settlements. https://www.bis.org/publ/bppdf/bispap125.htm

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences, 121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Romelli, D. (2025). Trends in central bank independence: A de jure perspective. In E. Farvaque & P. Stanek (Eds.), *Are central banks still conservative?* (pp. 59–95). Edward Elgar Publishing.

Contact

Martin Bok, Prague University of Economics and Business, W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic, bokm00@vse.cz; Dominik Stroukal, Metropolitan University Prague, Dubečská 900/10, 100 31 Prague 10, Czech Republic, dominik.stroukal@mup.cz