Balancing agency, transparency and accountability between human and AI agents: practical insights from the FinTech industry.

Wojtek Buczynski

University of Cambridge; London Business School; Fast Audit AI; FirmView AI; wb292@cam.ac.uk

Jingkun (Charly) Zhu

Kings' College London; Tsinghua University; Fast Audit AI; FirmView AI; charly.zhu@firmview.ai

Agentic AI appears to be a great fit for financial services, because it combines a degree of openness with structure; put simply, the very nature of agentic AI workflow very closely matches the nature and structure of many specialist roles within financial services, chief among them the analytical ones.

Using two real-life agentic AI use cases – financial audit and equity research – we will present our position on human-agentic AI interactions with emphasis on transparency, auditability, attribution and accountability. Our background is interdisciplinary as we are academic researchers as well as former financial services professionals and, currently, founders of two AI FinTech start-ups extensively leveraging agentic AI. Consequently, we represent the practical approach backed by academic background.

Our contributions to the workshop are:

1. In-depth discussion on the practical considerations of human-agentic AI interaction using two real-life use cases from a highly regulated industry.
2. Presentation of a number of potential "building blocks" for a robust and versatile human-agentic AI interaction framework that could be utilised in other use cases and industries.
3. A number of promising topics ("prompts") for future research.

CCS CONCEPTS • Human-centered computing → Human computer interaction (HCI) → HCI design and evaluation methods; Interactive systems and tools

KEYWORDS: Artificial Intelligence (AI), agentic AI, Large Language Models (LLMs), HCI, financial services, audit.

ACM REFERENCE FORMAT: In Proceedings of AutomationXP26 Workshop of the 2026 CHI Conference on Human Factors in Computing Systems, April 14, 2026, Barcelona, Spain. ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Financial services industry has long been considered a champion and early adopter of emerging technologies: mobile experiences, metaverse, mixed reality, quantum computing, and AI in all its varieties: from Machine Learning, through Large Language Models and now agentic AI.

One needs to be mindful that financial services remain one of the core and critical services of the modern world: they may not be as "existentially" critical as healthcare, transportation or defence, but they account for approximately 25% of the global economy – and they are critical to everyday functioning of businesses and individuals. In light of their significance and with the enormous repercussions of the 2008 financial crisis still fresh in memory, financial services are understandably a heavily regulated and scrutinised industry. Not unlike healthcare or air transportation, financial services are striving for a zero-tolerance margin for analytical errors as well as reputational and strategic judgement calls.

While agentic AI is a very recent technology, financial services had been pursuing various automation solutions for years before it emerged, for example (pre-AI) Robotic Process Automation (RPA). Agentic AI therefore does not necessarily introduce a paradigm shift as regards broadly understood "intelligent automation" – it takes an existing one to the next level. At the same time, the financial industry may not have yet reached the stage at which human-agent interactions are so established that they need to be rethought for the era of advanced agentic AI. We would argue that the financial industry first needs to design those on a foundational level from first principles. It is what our academic research and our AI FinTech startups are focusing on.

Interestingly, the highly regulated nature of the industry and its "non-negotiables" of transparency, attribution and accountability could be framed as a benefit rather than a challenge or an obstacle. Firstly, they set some clear rules for human-AI interaction and collaboration. The human-agentic AI relationship in financial services simply *cannot* be opaque or obscure. Secondly, they set a number of implicit safety limits for AI's agency and autonomy (for example, allowing AI agents to autonomously trade in financial markets would likely be considered an unacceptable risk, especially given well-documented automation failures in electronic trading such as the Knight Capital incident in 2012, where insufficient deployment controls contributed to severe market impact and

substantial losses[4]). Thirdly, solutions and frameworks meeting those strict standards are likely to meet standards of other industries too.

Finally, there is the key question of whether AI – agentic or otherwise – should be regarded as "hostile" competition or beneficial enhancement / augmentation by its human users. This is neither an academic nor a philosophical question – it's a question that directly informs the level of agency given to AI / retained by humans as well as the nature of human-AI interactions. We consider the antagonistic approach ("humans vs AI") as suboptimal to the collaborative one ("humans and AI"); we first wrote about it years ago[1] and we have only solidified our conviction since.

2 BALANCING AGENCY, TRANSPARENCY AND ACCOUNTABILITY BETWEEN HUMAN AND AI AGENTS

2.1 Financial audit

Audit, in its nature, is not exceptionally difficult intellectually. It is, however, exceptionally laborious, increasingly complex – and, importantly, it has not changed much in decades. While screens and spreadsheets have replaced paper financial statements, audit remains largely about looking at numbers, analysing them, and reconciling them. Over the years, as the size and complexity of business has been growing, so has complexity of the audits – as well as their costs, which have been rising at astonishing rates[3].

One of our startups – Fast Audit AI – is focused on a specific task called anomaly detection or in other words: finding numbers that do not fit. Anomaly detection is all about detecting patterns and then outliers to these patterns – it is a task Machine Learning seems ideally suited to; which it is. The core anomaly detection system is a combination of rules-based and Machine Learning algorithms.

The challenge is that identifying an outlier is of limited value to the auditors – without context, without explanation, without knowing "why is this an outlier?" there is little they can do with this insight. This is exactly where LLMs can add substantial value – they can provide context and / or offer some plausible, factual explanations as to why a given number is what it is – and particularly, whether the cause may be perfectly legitimate; or whether this is something that needs to be investigated thoroughly as potential wrongdoing. Unfortunately, in the current state of generative AI, there is no way to guarantee full factual accuracy in RAG (retrieval augmented generation) tasks. This creates a seemingly irreconcilable friction between the stakeholders: the auditors, the audit oversight bodies (Financial Reporting Council in the UK) and us, the AI vendors.

Firstly, we started by strategically selecting the LLM itself. We went for an open-source model to avoid dependencies on a third party (cost, opaqueness, behind-the-scenes model updates etc.). As the initial release of our system was targeted at British and European clients, we chose a European LLM for a better contextual fit. In the future, as we expand coverage, we are considering employing a US LLM for the American clients and a Chinese LLM for the Chinese ones.

Secondly, we extensively trained the model in-house – to the extent that we had the complete LLM running on our local machines. We fed the LLM with the curriculum of professional accountancy qualifications and then fed it thousands of financial statements, earnings call transcripts and other relevant inputs from the universe of the companies we cover.

Next, we moved away from a "standalone" LLM approach toward a combined LLM + agentic workflow. Consequently, we combined the LLM with an agentic orchestration layer: a task-level agent decomposes the user request, dispatches specialised nodes/tools (retrieval, reading/extraction, reconciliation checks), and decides the next step based on structured intermediate outputs. This setup produced immediate improvements in scope and depth by decomposing work into tool-using subtasks.

The system is coordinated by a Planner agent that produces an explicit, executable next-step decision, following the broader "plan–execute" and "planner–executor" separation that has been shown to improve controllability and efficiency in tool-augmented LLM systems. With LLM as agentic AI's orchestrator, evaluator and supervisor we discovered a simple way to manage a number of AI agents within a single, unified process.

In high-stakes finance workflows with little to no margin for error, we do not rely on the agent alone to control hallucinations, because small factual errors can compound across steps. We therefore adopt a tiered review policy: (i) for numeric facts and extracted financial data (e.g., values, units/currency, periods), human verification is mandatory; (ii) for templated narrative completion where the outline and evidence are fixed, we use spot checks; and (iii) for consequential analytic claims, we require full human read-through. To reduce reviewer burden, we additionally use a cheaper verifier model to flag inconsistencies for human escalation.

The entire system is "packaged" into a user-facing chatbot we called Charly – "Charly the chatbot" aka CtC – which was added to the existing, Machine Learning-based anomaly detection system. This made the ultimate synergy: quantitative analysis of financial statements augmented by qualitative insights from CtC. CtC is designed not to give definitive statements (something conventional LLMs are very good at, even if they're confabulating) and to speak in terms of what could be the likely explanation. Furthermore, we

tried to make it so that whenever possible CtC presents a range of plausible explanations. This "inconclusivity" is important when auditors are the key stakeholders. Auditors need to retain their intellectual autonomy and professional judgment – fundamental tenets of their industry – with our solution positioned as an analytical support tool.

Lastly, there are two important outcomes from the perspective of attribution and transparency:

i. We record trace for each user request: the Planner's step decisions, tool invocations, intermediate structured outputs, and the evidence objects used downstream. This enables targeted debugging and iterative system improvement.
ii. We make contribution and responsibility explicit at the artifact level. Each intermediate result that can affect the final answer is captured as a small evidence object with provenance (source/type/identifier, retrieval parameters, timestamp) and a pointer to the producing module/version. The user-facing response is produced only after a final synthesis step and is released subject to verification gates. In other words, specialist agents can contribute evidence, but publication authority is constrained by orchestration and human oversight, which is essential in audit workflows.

2.2 Equity research

Equity research is a fundamental part of the investment industry – it is also tightly regulated. It serves two main purposes:

i. It provides a snapshot of the current situation of the covered company including all material information, developments and financial performance, prepared by an expert analyst.
ii. It provides share price forecast for a fixed time horizon (usually 6 – 12 months) along with a simple recommendation: Buy, Sell or Hold.

Typically, a research report is a combination of free-form text and financial data. It is the analyst's personal judgment to decide what information they consider material and relevant. The price forecast is the ultimate expression of the analyst's judgement and should be a logical conclusion of the facts, numbers and opinions included in the report.

In our equity analysis startup – FirmView AI – we found four practical use cases for agentic LLMs in investment workflows: (1) automated news aggregation on companies and geopolitical context; (2) automated aggregation of user/consumer reviews and commentary; (3) semi-automated financial data collection from trusted sources; and (4) semi-automated investment research generation. In this submission we focus on (3) and (4), because they demand the strongest auditability, provenance, and human review policies. (3) and (4) are perfect use cases to utilise LLM + agentic AI combo. Information that can be material and have impact on a company's performance, and, by extension, on its share price is vast, unbounded and heterogenous. It could be general economic and market developments (global or local); it could be industry-specific impacts; it could be political or regulatory developments; lastly it could be random, unforeseeable events like Covid. While research reports and financial statements tend to be highly structured and orderly, the starting point is subjective information sourcing, data selection and filtering. Historically, this has always been a domain of human analysts – highly-paid professionals who honed their talents and instincts through years of practice. As the job has a degree of subjectivity and judgement call, it has been considered highly "cerebral" and generally safe from AI disruption. Then came LLMs, followed by agentic AI, and they started to change everything.

Evidence scope and release controls are central in this domain. In our current production workflow, externally verifiable evidence for research reports is restricted to primary sources: company filings/financial statements and the covered company's official website. Financial data collection (use case 3) produces candidate evidence items that require human evaluation before they are admitted into our internal evidence store; research generation (use case 4) is then constrained to cite and compose only from these approved evidence items, resorting to targeted web retrieval only to fill clearly identified gaps within the same primary-source scope.

LLMs can produce professional-looking equity research, but without strict evidence control such reports risk being factually incorrect (hallucinated) or otherwise useless ("AI slop"). With all the data available online it is possible to arrive at new, original insights – so the tool is not misplaced, and the use case does make sense; it just needs very thoughtful workflow design. In our approach workflow design – evidence capture, conflict handling, review gates – is the core consideration. To operationalize attribution, we represent evidence as atomic units that store (i) the extracted snippet/value, (ii) provenance metadata (source identifier, retrieval query/parameters, timestamp), (iii) the producing module/version and an execution-trace pointer, and (iv) optional typed links to other packs (e.g., supports/contradicts) when reconciling conflicts. This design makes multi-agent contributions legible to reviewers and enables both automated consistency checks and efficient human verification.

Our approach to developing successful human–LLM collaboration is guided by the following tenets: (i) clearly defining the use case and decomposing it into tasks (data acquisition, synthesis, report drafting), then into smaller self-contained subtasks; (ii) selecting appropriate, task-specific technologies for each subtask (tools, retrieval, numeric checks, structured extraction, constrained writing); (iii) defining stakeholders involved and their needs, objectives, and constraints; and (iv) designing robust workflows with technical

quality controls and humans-in-the-loop, prioritizing transparency and auditability: reducing unsupported claims (hallucinations) during generation and enforcing explicit verification and expert review before release.

3 CONCLUSIONS

Looking at the two use cases above – financial audit and equity research – we can see that one thing that has a fundamental impact on the entire AI workflow and the human-AI interaction is how the human end-user consumes the system's output:

- If it's a live interaction (a chat with CtC, similar to a chat with ChatGPT) then there is some degree of risk of hallucinations because Fast Audit AI can only review and fine-tune the system ex post. This is something our clients need to be mindful of; a degree of (currently) irreducible risk.
- If it is consumption of static, "pre-packaged" content (as the case is with FirmView AI's equity research reports) then it is possible to add human analysts at the very end of the report production chain (in addition to them being included in earlier stages as well).

In other words, whether the user interacts with a live system (chat-like) versus a static artifact (a report) fundamentally changes the risk profile and control strategy. Static report generation enables explicit verification gates and a final expert review before release, which in turn makes it feasible to reduce unsupported claims below the workflow's release threshold.

From our industry observations the impression is that it is universally assumed that human-agentic AI interaction is always going to be live (to clarify: we mean interaction between agentic AI and external users / clients; in-house interactions do not count because the risks remain contained). We would like to challenge this assumption and initiate a discussion as to whether this is really the case. Having discussed AI agents in great detail above, we would now like to focus on their human counterparts. Looking from the business perspective questions about human agency oftentimes seem like euphemisms for questions about human (un)employment. In financial services – and in business in general – the bottom line is usually money. It may seem crude, but it helps (re)frame the discussion.

The question about the level of human involvement – in other words, about the optimal trade-off between human and AI agency – is fundamental to our discussion. If (when) agents augment human performance then fewer people may be needed; case in point: supermarkets self-checkouts – human cashiers still exist, but there are far fewer of them than a decade or two ago. Same may be the case for modern-day finance professionals, such as auditors or analysts. However, our in-depth look at two highly "AI-augmentable" professions above – auditors and analysts – points to more nuanced and encouraging conclusions. Holding the expectation of high quality of output constant – which is obvious and non-negotiable across all industries – there is a certain limit (a "floor" in financial parlance) on the level of human involvement; particularly in cognitive jobs. Furthermore, agentic AI could once again be seen as an opportunity rather than a threat: instead of fewer people producing the same amount of output of the same quality as pre-agentic AI the same number of people could produce more work, and / or of higher quality than before. This could lead to competitive advantage, and financial services – as well as, presumably, most other industries – are extremely competitive. Plus, with AI shifting certain quality expectations upwards for the entire industry, the pre-agentic AI baseline of performance quality will likely no longer hold in the new agentic AI paradigm; it will shift upwards, which will be all the more reason to keep human involvement at a reasonable (more than minimal) level.

At the same time, we think that agentic AI "solving complex tasks with minimal human input" – as the workshop brief put it – may be a slight overstatement. There are a number of complex tasks agentic AI should be able to solve in theory – but our experience as practitioners showed us that currently there is still a gap between theory and practice. This gap, in our view, explains what we consider fairly limited progress in the implementation of AI to date. It could be argued that the next generations of agentic AI – and other types of AI that haven't broken into mainstream yet – will bridge that gap. We have no way of knowing that, but autonomous self-driving robotaxis serve as a profound example in how difficult it is to cross from "working in controlled, artificial conditions" to "working (almost) autonomously in the real world". In financial services there is also the regulatory onus. The downside is so serious that retaining humans and their agency is likely to remain a reasonable choice even if they can be fully replaced. Not all industries are as tightly regulated as financial services, but most, if not all of them, face reputational, operational and financial downside should autonomous technology go wrong without "humans in the loop". Consequently, we prefer to frame agentic AI as augmenting and enhancing human agency (or, to put it in business terms: performance); not threatening it. We see the real challenge elsewhere: it is in organisational structures, hierarchies and behaviours that oftentimes require substantial re-engineering to accommodate new technologies and to truly enable synergies of human-AI interaction. Unfortunately, these structures are often deeply entrenched, with many people invested in defending them.

Sustaining human skills is a profound consideration. Even though we expressed some scepticism regarding the extent to which "agentic AI handles complex problem-solving with minimal human involvement" (to quote the workshop brief) in real-world

professional environments, the question of skills remains a valid one. Traditionally, across skilled and especially highly-skilled professions – such as auditors, analysts and countless other specialisations within financial and professional services – skills were acquired and honed first through university studies and then, for the entirety of one's career, through acquired experience. Agentic AI handling tasks that would previously be done by humans – particularly the simpler, more repetitive and rote tasks traditionally given to apprentices, graduates and other junior staff – disrupts the traditional model of a professional career. We are explicitly mentioning junior employees because their tasks are more likely to be taken over by agentic AI than the more complex tasks of their more experienced colleagues; the problem is that if there are disproportionately fewer juniors, then where will tomorrow's seniors come from? At the same time, it needs to be stressed that for-profit businesses' objective is not sustaining human skills – their objective is making money. Education, skills, employment – those are societal and political considerations.

Human – AI interaction is going to be hugely impacted, informed and driven by user interfaces mediating this interaction. In our observation it is currently a somewhat under-appreciated consideration both in tech industry as well as in professional services. Our - Fast Audit's and FirmView's – systems' user interfaces need to be clear and transparent enough to convey the value of our analytics (or, in other words: to persuade the clients to buy our products). Our guiding principle has been a balance of transparency, in-depth insights and focus on the essentials. We discovered during the design process that clients appreciate having all the transparency disclosures available – but not necessarily in their face. Consequently, we wrote very thoughtful and comprehensive clarifications and disclaimers, which users are never further than two clicks from. We do not overwhelm them with AI transparency disclosures within core analytical functionality of our systems. Our internal interfaces are minimal (some of them resemble the old MS-DOS command prompt) and designed with transparency and auditability as guiding principles. *All* actions and reasoning chains of AI agents are fully logged; in fact, the agents themselves are designed to be very "extroverted". The internal users are our engineers who expect minimum aesthetic distraction and maximum transparency, audit trails and insights they can turn into system enhancements.

One of the open questions raised by the workshop's organisers is "when does additional transparency become counterproductive?". From the perspective of financial services (and their regulators) the answer is most likely going to be: never. There is no such thing as too much transparency in finance and we cannot envision this is likely to change in foreseeable future. We cannot speak for other industries, but in our interdisciplinary experience as AI researchers and engineers we have not yet encountered a situation where any amount of (additional) transparency could be in any way counterproductive or otherwise problematic. If anything, we would expect that as AI becomes ubiquitous and increasingly complex as well as embedded into critical systems, the regulators across industries will be elevating their requirements as regards transparency and auditability accordingly. At the same time, various stakeholders may require different types and degrees of transparency. We therefore adopt progressive disclosure: the primary interface surfaces conclusions with a compact evidence summary, while full provenance and step-level traces remain available within a small number of interactions (e.g., 'two clicks away') for users who need to audit, contest, or debug the result.

Separately, regulatory considerations need to be factored in as one of the abovementioned "non-negotiables". Financial services are tightly and explicitly regulated, and those regulations have direct bearing on how humans interact with agentic AI, and the levels of agency given to AI / retained by humans. Many European enterprises – for-profit and public services – will to some extent fall under the aegis of the EU AI Act[2]. Even if they do not fall under the category of explicitly regulated high-risk AI systems, they will fall under the "catch-all" Article 95, which covers *"drawing up of codes of conduct, including related governance mechanisms, intended to foster the voluntary application to AI systems, other than high-risk AI systems, of some or all of the requirements set out in Chapter III, Section 2 taking into account the available technical solutions and industry best practices allowing for the application of such requirements"*. We interpret Article 95 as a *de facto* requirement to implement most provisions of the act in a flexible, proportionate manner – with national and sectoral regulators deciding whether individual firms have implemented those – technically non-binding – best practices to their satisfaction.

The two in-depth use cases we presented above have taught us a number of practical lessons on human-agentic AI interaction which we believe are universal and "generalise-able" enough to be of benefit to the workshop participants regardless of what industry or sector they work in. We think that the fact that we implemented them in real-world, client-facing solutions made our insights grounded and practical. We presented a number of considerations that can be used as "building blocks" for human-agentic AI interaction frameworks (delineating between live and static interactions, discussing "humans in the loop" at different stages of the workflows, auditability and transparency, regulatory considerations) across various industries, both for-profit and otherwise. Lastly, much of our work is informed by present-day cutting-edge research and is a "prompt" for future research across agentic technology per se as well as human-agentic interaction, human agency, determining minimum safe level of human involvement and oversight, and many more. We look forward to continuing our contributions to this fascinating area both on practical (business) as well as academic fronts and hope to be able to share our insights with the workshop participants.

REFERENCES

[1] Wojtek Buczynski, Fabio Cuzzolin, and Barbara Sahakian. 2021. A review of machine learning experiments in equity investment decision making: why most published research findings do not live up to their promise in real life. (2021); International Journal of Data Science and Analytics. *11*, 221–242. https://rdcu.be/ch7Xo.

[2] European Parliament and of the Council. 2024. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 12 (Record-keeping). Official Journal of the European Union, L, 2024/1689 (12 July 2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[3] Ideagen. 2023. *audit fees report (2023)*. Retrieved from https://staging-corp-waf.ideagen.com/thought-leadership/whitepapers/2023-audit-fees-report.

[4] Steve Schaefer. 2012. Knight Capital Trading Disaster Carries $440 Million Price Tag. *Forbes*. https://www.forbes.com/sites/steveschaefer/2012/08/02/knight-capital-trading-disaster-carries-440-million-price-tag/