The Economy Research

THE ECONOMY / RESEARCH PAPER SERIES

NOVEMBER 2026

[AI and Cyber-Security] The Economics of Delegating Authority to AI

Procurement, Review Capacity and Break-Even Controls

The Economy Research Editorial
The Economy Research

RESEARCH PAPER

economy.ac/research

About the paper

This paper examines the economic benefits and risks of delegating business processes and payment authority to AI agents, distinguishing model capability from enterprise access, policy controls and the human cost of review.

Reading the evidence

It combines reported observations about payment fraud and corporate controls with simulations of agent behavior and scenario valuations. All modeled behavioral rates and financial results rely on stated assumptions and are not empirical forecasts.

Author

The Economy Research Editorial
The Economy Research

71 Lower Baggot Street
Dublin 2
Co. Dublin, D02 P593, Ireland

Publication

The Economy Research
23 November 2026
economy.ac/research

Originally published
23 November 2026
The Economy Research Editorial
Read the original paper

Contents

Abstract 1

1 Introduction - Delegated Authority Is a Procurement Decision as Much as a Model Choice . . . 1

2 Separating Model Capability from Enterprise Access 3

3 Who Authorizes Actions, Owns User IDs, and Restores Failed Processes 5

4 Net Benefit after Integration, Review, and Residual Loss 9

5 Where a Stronger Control Pays for Itself 12

6 Conclusion 15

References 16

Appendix: Scenario Values 17

Eight figures and three tables accompany the analysis. The reference list identifies the works cited in the manuscript.

Abstract

While an AI-powered invoice-to-payment system will typically be procured on the basis of its model performance, its actual net benefit will largely depend on the user ID assigned and the pre-action controls applied. This paper explores the extent of control an entity may wish to grant to such a system and the associated net benefits. It derives this from behavioral rates obtained from a parallel SIAI study, considering 4 scenarios: wide and narrow user IDs with and without a fixed pre-execution control, for a postulated model adherence to malicious prompts. These are then applied across 4 processes: receiving documents, updating customers' records and making standard and material supplier payments. Staff pay rates are taken from publicly available U.S. statistics; the cost of integration is assumed and time per review, malicious prompt injections and unrecovered losses are varied. For standard supplier payments, removing all pre-execution controls and granting the highest possible level of authority results in ~$194k p.a. net loss. Granting a lower level of authority and applying a pre-execution control yields a net gain of ~$138k p.a., driven primarily by reduced error in standard payments, with attack mitigation a secondary factor. With a pre-execution control, the incremental net benefit of increasing the model's ability drops to ~$8k p.a. A narrower user ID is most advantageous where information could flow externally and a pre-execution control where funds are transferred. Under the base case, there is no point in automating receipt of documents or processing a low number of material payments, as the cost of integration cannot be recovered, irrespective of the scenario. Approvers rarely hold real control, with the teams responsible for establishing and maintaining access, setting and amending controls, identifying outliers and bearing residual loss being distributed elsewhere. All behavioral rates and monetary figures used here are assumed. However, the relative position of each scenario is likely robust to these.

1 Introduction - Delegated Authority Is a Procurement Decision as Much as a Model Choice

A company receives around 4,000 invoices from suppliers each month. It is therefore interested in learning more about an AI system which will examine all incoming invoices, compare them with relevant information on other documents and propose an appropriate payment. A demonstration by the vendor will establish broadly the range of functions which the model will perform satisfactorily. Many further issues arise, all of more limited scope, which need consideration and resolution. Will the system propose payments for approval and release for execution, or merely complete a payment request? Should it amend the bank account details of the supplier if notified in the invoice? Should payments up to a given limit be executed without review? With what user ID and associated permissions would the system act? There will be many such questions, each transferring some function and the associated authority currently exercised by staff in accounts payable, their supervisors, or in treasury. Each will have some combination of costs, in the time required to review the system's proposals, development time

and the risk and magnitude of error. It is unlikely that a vendor's demonstration would give much indication of any of these.

The majority of organizations will be subject to frequent attempts at fraudulent payment. 76 percent of US companies surveyed by the AFP in 2026 suffered one or more attempts to obtain fraudulent payments during 2025 and 74 percent one or more instances of business email compromise, or BEC, compared with 63 percent during 2024.[1] Controls were very widely adopted: requiring approval by a senior level of management for payments over a given limit, by 94 percent of respondents; making an independent check with a known phone number of any instruction to make or amend a payment, by 94 percent; and having procedures to check any change of details on invoices and of bank account and other contact details, by 96 percent. Attempts to obtain fraudulent payment were detected by staff in treasury in 83 percent of cases and successful frauds in 55 percent.[2] All these measures presume human intervention to read communications, identify inappropriate changes to details and make a telephone call. An agent deployed in this area would presumably receive all relevant communications; it would not be expected to spontaneously make a telephone call to check information and so the question becomes how many and which, of the existing measures would continue to apply and whether there might be more cost-effective alternatives.

Accordingly, this paper considers how much function and so authority, might profitably be delegated to an agent and the net benefits. As indicated above, the companion paper considers some of the relevant behavior, simulating the outcome given an assumed frequency with which the model will seek to execute inappropriate actions. There are four cases considered, according to the breadth of permission given by the user ID used by the agent and whether there is further deterministic testing of all actions with effect on internal records or external communication, applied as the action is executed.[3] All assume the same underlying model and processes and the same rules as to the extent of authorization for the tasks set. It assumes that the model will seek to follow malicious instructions included in the input documents in 18 or 35 percent of cases, according to the capability level. These high rates of response to attack are assumed, based on published tests and not further investigated. This paper will calculate relevant benefits, costs of compute resources, development and integration, review of proposals and expected loss from errors and failure, across 4 types of tasks or workflows. Published rates of pay, or other assumed values, are used. It is therefore possible to identify the maximum acceptable cost of additional measures, that is, the cost at which the marginal benefit of tighter control is exhausted.

Chapter 2 summarizes the considerations in procuring the capability of a model and allowing it access to internal processes and records of an enterprise, with a summary of results from the companion paper. Chapter 3 reviews some aspects of appropriate allocation of authority, in particular ownership of the user IDs used and responsibility in case of failure, with reference to published information on fraudulent payments in the U.S., UK and Eurozone. Chapters 4 and 5 describe calculations of net benefits of different configurations and break-even levels of control, respectively.

2 Separating Model Capability from Enterprise Access

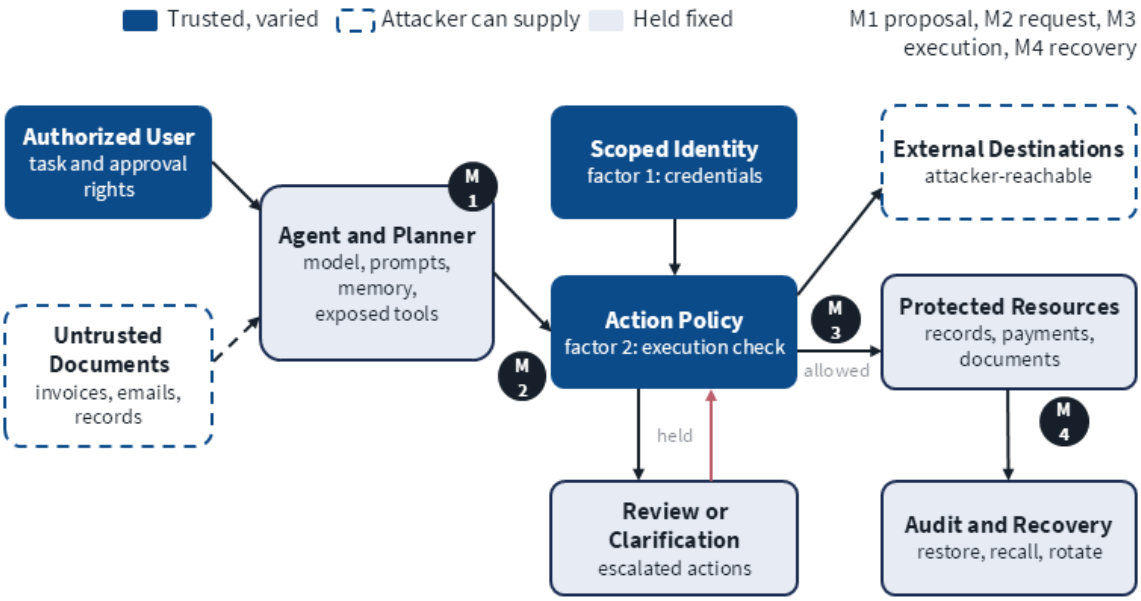
It appears normal for an enterprise to procure these two elements separately. It will evaluate and purchase a model, judging its suitability by published and other test results and pay according to usage. It will then create one or more user IDs and grant such access to various services and other facilities, with a duration of authorization, as is considered appropriate, normally as a function performed by staff responsible for the relevant directory or similar facility. In May 2025 Microsoft announced the provision of user IDs for use by AI agents created in Copilot Studio and Azure AI Foundry, to be added to its directory product, Entra. Agents would be authorized to request time-limited authorizations appropriate to required resources.[4] It appears that the provision of appropriate access to enterprise facilities for AI agents is now offered as a significant range of products. OWASP includes 'excessive agency' in the Top 10 for LLM Applications for 2025, which can result from excessive functionality, authorization and degree of autonomy permitted. All are features of installation and configuration, not the model.[5]

It would appear that agents have been widely procured, but controls lag. SailPoint surveyed staff with security and other IT responsibilities and, in May 2025, reported that 82 percent of respondents' organizations used AI agents, but only 44 percent had policies addressing their security. Agents had performed unintended or undesirable actions for 80 percent of respondents; in 39 percent they had accessed unauthorized systems or resources and downloaded confidential information in 32 percent. In 23 percent of organizations, an agent had been induced to disclose authorization information.[6] There is no information from this source as to the frequency of such occurrences. The significant point is that these incidents are all associated with the extent of access permitted to agents, unlikely to be a major consideration in deciding whether and what model to procure. Consistently, authors contributing to Brookings in September 2026 identified various controls applied during operation of the system by the customer as the most likely effective response to agent risk.[7]

The companion paper therefore considered the effect of restricting the range of assets accessible by an agent's user ID to those relevant to current tasks and imposing a policy check of authorization applicable to each task prior to the execution of all actions changing state or communicating externally. Agents were deployed in simulated simple workflows to process documents, update records relating to customers and make payments to suppliers. It assumed rates of compliance with injected malicious instructions as described in §1, 18 and 35 percent. The proportion of all malicious attempts successfully executed ranged from 31.9 percent with maximum authorization and no policy check, to 8.1 percent with restricted user ID and policy check, with template-clustered confidence intervals of 26.7-37.2 and 5.3-10.8 percent, respectively. Successful completion of tasks when subject to attack increased from 72.5 to 79.4 percent and was hardly affected when there was no attack.[8]

Figure 1. Credentials and the execution check act at different points between a proposal and an executed action.

Workflow components by trust status and measurement point



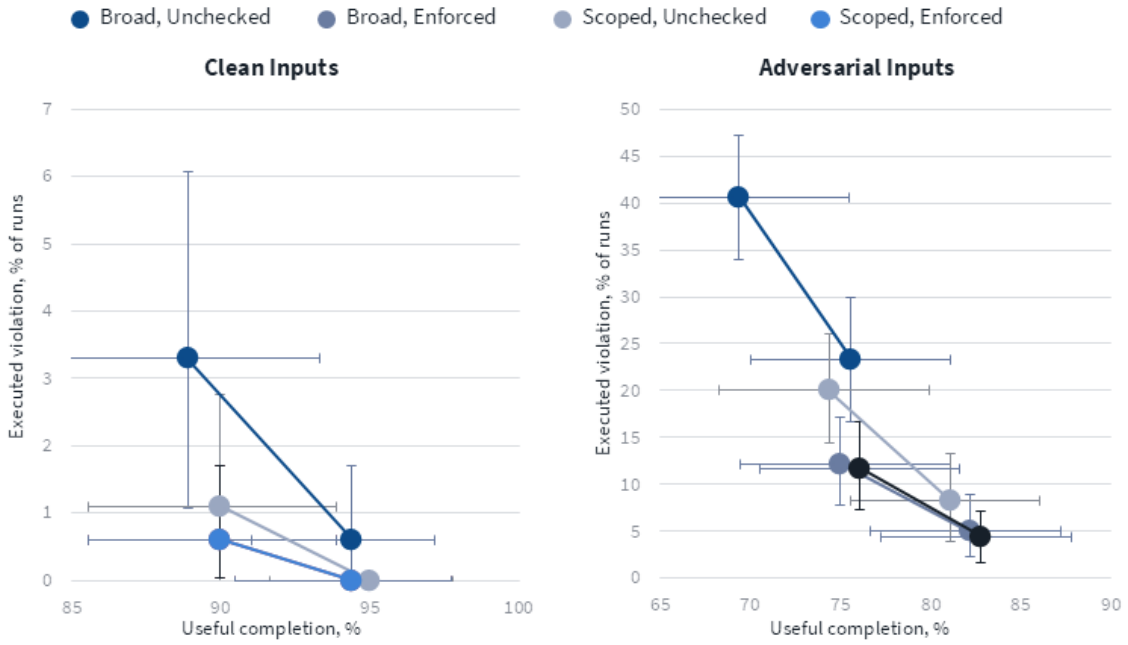
Note: Scope sets what the identity can reach; the policy checks each state-changing or outbound action before it executes. M1 to M4 mark where proposals, requests, executions and recovery are measured.

Source: SIAI Research Editorial (2026) [AI and Cyber-Security] Agent-level versus system-level safety: permission scope, action enforcement and workflow composition. SIAI Research working paper. Swiss Institute of Artificial Intelligence.

There was a substantial common effect of the two types of control: policy check reduced successful attacks by 23.3 percentage points without restricted authorization, compared with only 6.1 points when authorization was restricted. In the largest set of simulations in the paper, 96.3 percent of all successful attacks with both types of control took place within authorized functions, for example, paying excessive amounts to a valid supplier; they would be indistinguishable to the policy from legitimate actions. Although 96.3 percent of all successful attacks, compared with 25.8 percent with maximum authorization and no policy check, occurred within the authorized scope of the task, they would generally be detected only on subsequent reconciliation and the median time to return records to an appropriate state was 45 hours compared with 32 hours. These checks would be expected to incur no additional cost of model invocations. They would result in some requests for approval: 21.4, 7.7 and 26.7 per 100 runs with policy check only, restricted user ID only and both, against no requests without either.[9]

Figure 2. Under attack, the controls cut executed violations and raise useful completion.

Completion and executed violations, % of runs



Note: Model simulation under stated assumptions. Each line joins the lower- and higher-capability model profiles; bars are 95% intervals from a template-clustered bootstrap of the 2,880-run design.

Source: SIAI Research Editorial (2026) [AI and Cyber-Security] Agent-level versus system-level safety: permission scope, action enforcement and workflow composition. SIAI Research working paper. Swiss Institute of Artificial Intelligence.

Applying assumptions in Chapter 4 below to these outcomes, it is apparent that the choice of user ID authorization and policy check has a much greater effect on the net benefits of typical payments than does the level of model capability. Without policy check, deployment of the higher level of model capability would be expected to be worth around $183,000 per annum more, largely through reduced numbers of payments to inappropriate suppliers. However, once there is a policy check at the point of execution of an action, the difference between model capabilities reduces to around $8,000, while the introduction of a policy check would increase net benefits by $301,000 to $476,000 per annum depending on model. It would therefore be unreasonable for an organization to devote substantially greater effort and time to evaluating models than to determining an appropriate user ID for deployment, although this conclusion depends on the assumed error rates and frequency of proposed inappropriate action.

3 Who Authorizes Actions, Owns User IDs, and Restores Failed Processes

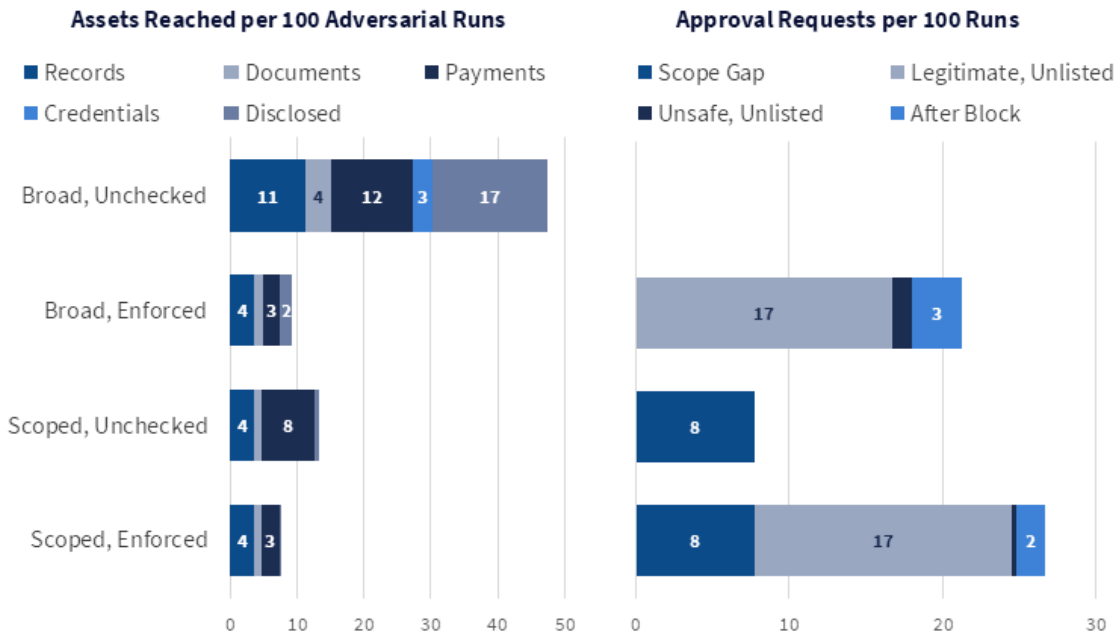
Aghion and Tirole define and distinguish between formal and real authority. The former may be thought to be authorization to make a decision, while the latter is the practical ability to do so. They demonstrate various cases in which the formally authorized party will prefer to approve decisions made by another party with superior information, so that real authority lies with that party.[10] Many cases of approval of an action proposed by an AI system would

appear to fall into this category. For example, staff asked to approve a payment made on the basis of an invoice will normally see only the name of the payee and the amount and so have no information as to any instruction in the invoice itself and little else. Approval will transfer formal authority for the payment to the approver, but real authority remains with the model and author of any instruction.

Garicano provides a simple model of an organizational hierarchy, in which problems encountered at lower levels and not resolvable with available information and skills will be passed to progressively higher levels of staff with different and more valuable information and skills. The extent of the hierarchy will depend on the proportion of production passed upwards at each stage.[11] Staff approval of actions proposed by an AI system subject to policy check would be analogous and one would expect the majority of actions referred for approval to be legitimate, failing to meet some rule of the policy, rather than malicious. In fact, 16.8 of 26.7 requests per 100 runs with restricted authorization and policy check were legitimate actions. For 48,000 invoices per annum, this would give around 12,600 requests for approval requiring, on average, 10 minutes, so 2,100 hours per annum. This exceeds the available hours of one staff member and would be a substantial burden on any small group. This volume would depend primarily on the ability of the policy to encompass normal transactions and relatively little on the incidence of attack.

Figure 3. Most review requests come from legitimate work that the policy does not list.

Assets reached and approval requests, per 100 runs



Note: Model simulation under stated assumptions. Assets are counted in adversarial runs; approval requests pool clean and adversarial runs. Values below 1.5 are not labelled.

Source: SIAI Research Editorial (2026) [AI and Cyber-Security] Agent-level versus system-level safety: permission scope, action enforcement and workflow composition. SIAI Research working paper. Swiss Institute of Artificial Intelligence.

Published guidance on appropriate internal controls generally advocates division of relevant functions. In particular, the COSO guidance of 2013 includes segregation of duties as one aspect of control activities, that is, separation of authorization, operation or recording and asset

custody.[12] The IIA's "three lines" model of 2020 allocates overall responsibility to management, appropriate controls and other functions as a second line and assessment by internal audit as a third line.[13] It is clear that an agent examining an invoice, updating records and instigating payment would combine several functions. It is therefore necessary to consider appropriate segregation around the functions performed by an agent. A possible allocation in a finance function would be as follows, although many variations are possible:

Table 1. Allocation of Functions and Responsibilities for Enterprise AI Agents

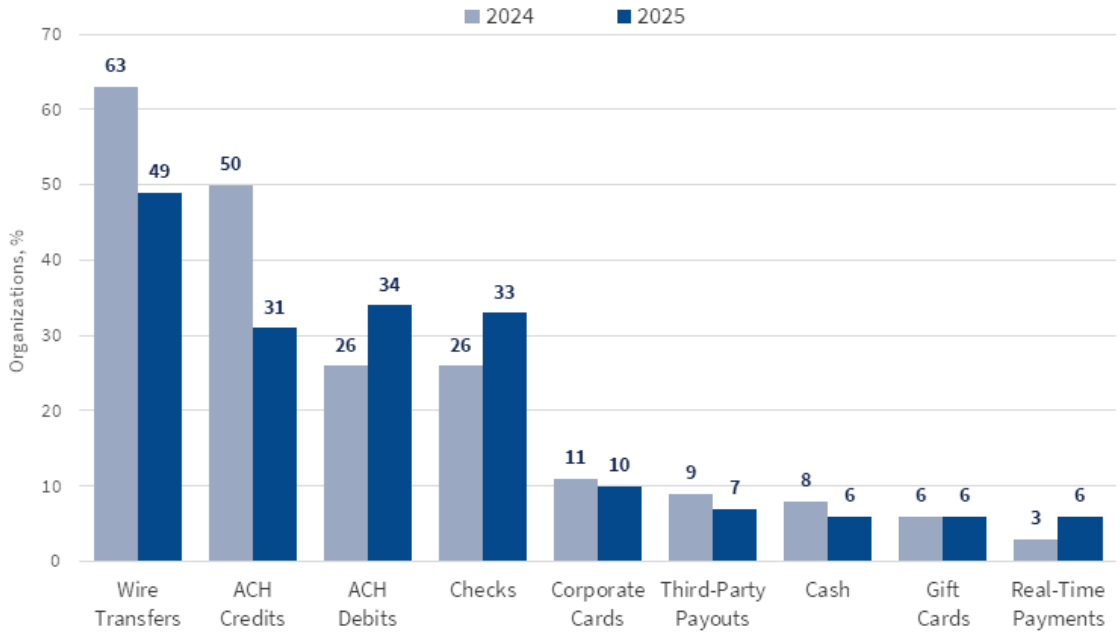
Function	Likely responsible	Relevant available information	Responsible for consequences of failure
select and procure model	IT function, sponsored by affected business area	various measures of model performance and unit price	Business area, additional costs of correcting errors, direct loss
determine authorization of agent's user ID	responsible IT function	range of authorization, periods of authorization and usage	Business area responsible for affected processes and records
define policy test and parameters	Second line function	rules applied and resulting referrals for approval	Staff required to approve proposals
approve actions referred	Management of area, for example, AP, treasury	Information provided by agent, unlikely to include original document	Company
correct consequences of failure	For example, AP, treasury, bank	Identification on reconciliation or otherwise, success of recalls	Company, net of sums recovered

It can be seen that none has all the information necessary to determine an optimum level of the function for which it is responsible: for example, IT setting authorization has little knowledge of records and processes that might be affected; those defining policy are removed from the burden of considering resulting approvals. It is to be expected, therefore, that each would tend to determine a level minimizing cost to themselves. So, for example, IT would tend to increase authorization of an agent where standard templates of action are interrupted, which occurred in 5 out of 60 cases in simulations. The second line would prefer conservative rules resulting in referrals for approval by others. It seems necessary for all relevant costs to be considered by one function, responsible for the process overall.

The methods of obtaining fraudulent payments continue to vary. Within the population of companies affected by BEC in the AFP survey, 63 percent in 2024 indicated that the method included wire transfer, falling to 49 percent in 2025; ACH credit from 50 to 31 percent; in contrast, ACH debit increased from 26 to 34 percent, cheque from 26 to 33 percent and real-time payment from 3 to 6 percent.[14] Agents instigating payment would presumably be capable of using several such methods and so would be subject to all such attacks. A check of payee details appropriate for one type of transfer would be irrelevant to another.

Figure 4. Business email compromise moved from wires and ACH credits toward debits, checks and real-time payments.

Payment methods used in business email compromise, % of organizations



Note: Respondents could name more than one method.

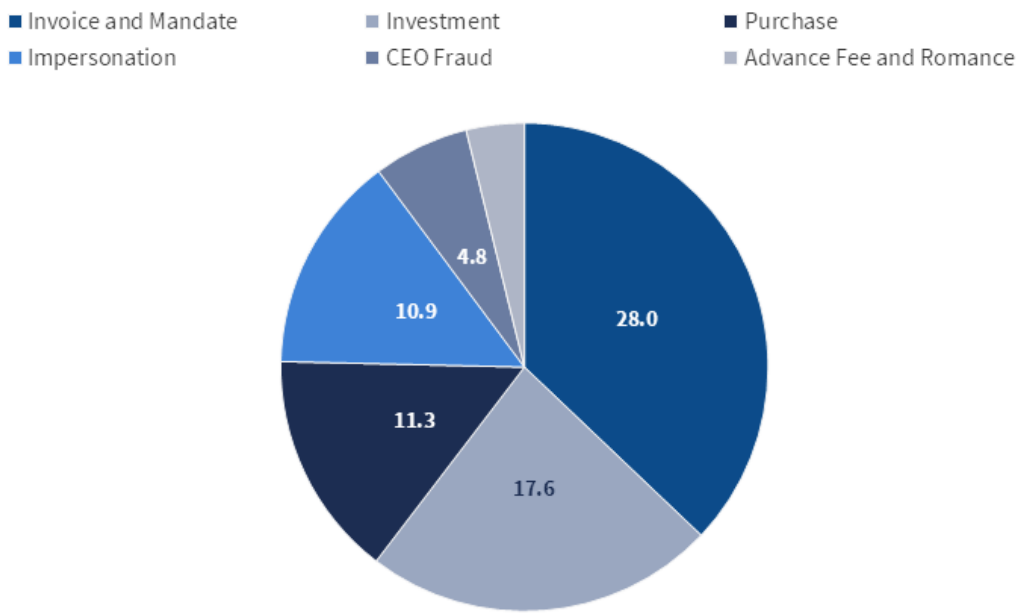
Source: Association for Financial Professionals (2026) 2026 AFP Payments Fraud and Control Survey Report: Key Highlights. Rockville, MD: AFP.

Prompt action is important if payment is to be recovered. The FBI reports total losses from BEC of $3.05 billion in 2025. Its Recovery Asset Team requested banks to stop payment on 3,900 occasions, representing attempted fraud of $1.16 billion, of which $679 million was stopped.[15] Consistently, the companion paper assumes the probability of successful recall falls as time from payment increases. Given that most successful attacks, continuing after application of policy checks and restricted authorization, would be within the scope of the task authorized, these will generally be the latest detected. It follows that the effect of controls will be substantially fewer losses, but slower recovery when loss occurs.

Various rules apply to the responsibility of others for unrecovered fraudulent payments. UK Finance reports total losses from authorized push payment scams from non-personal accounts of £75.6 million in 7,504 cases in 2025. Invoice and mandate fraud, that is, seeking to cause a legitimate payment to a supplier to be made to another account controlled by the fraudster, accounts for £28.0 million. Total sums repaid to non-personal victims were £34.6 million.[16] Scam victims paying up to £85,000 from personal, micro-business and charities' accounts are eligible for mandatory repayment under current regulation; so most other companies would be reliant on voluntary repayment by relevant banks.[17]

Figure 5. Invoice and mandate fraud is the largest source of UK business-account scam losses.

UK business-account scam losses by type, £ million, 2025



Note: Total £75.6 million across 7,504 cases. Impersonation combines police or bank staff and other impersonation; advance fee and romance scams, £2.9 million, are combined and not labelled.
Source: UK Finance (2026) Annual Fraud Report 2026. London: UK Finance.

There are some checks which will be made by others. All Payment Service Providers in the Eurozone must, from 9 October 2025, make available a service to verify the payee prior to authorizing a payment by credit transfer. The name provided by the payer is compared with the name or names associated with the account number provided.[18] So this would be effective against payment to a false account following amendment of records, which would be central to many instances of fraud by false invoice and mandate. It would be ineffective against an excessive amount paid to the correct account. If such services were provided centrally, for example, by a shared services organization, further issues would arise as to appropriate action if that facility failed, whether to continue with no check or suspend payments and responsibility for determining this.

4 Net Benefit after Integration, Review, and Residual Loss

The total expected net benefit of deployment of an agent in configuration a is assumed to be over 1 year, with development costs spread over 3 years, so:

$$V(a) = B(a) - C_{\text{compute}}(a) - C_{\text{integration}}(a) - C_{\text{review}}(a) - \mathbb{E}\big[L(a)\big] \tag{1}$$

Benefits are calculated only for tasks completed to the required standard, taking the time taken by manual performance. Failed tasks would be performed manually, giving no net benefit. Cost of compute resources is assumed to be $0.02 per invocation. Costs of development and integration are spread over 3 years, with ongoing maintenance added each year. A general amount is assumed to be required to integrate the model in the relevant processes and separate amounts to set up and maintain user ID authorization, to develop and maintain policy and optionally to improve the policy. Cost of review is total time spent in approving actions proposed by the agent multiplied by appropriate rates of pay. Expected loss is the sum over all relevant assets of value assumed to be lost if left unremedied, distinguishing between cases where there is an injected instruction and other error and time required to restore remaining assets to appropriate state. So time spent on restoration is included only in expected loss.

Rates of pay are based on published figures. The median wage was $24.36 per hour for bookkeeping, accounting and auditing clerks and $40.23 for accountants and auditors, May 2025.[19] Wages and salaries represented 70.0 percent of total compensation cost in the private sector, June 2026.[20] So, total cost per hour is assumed to be $34.80 for a clerk and $57.47 for an accountant. Other values assumed: time taken for manual completion of tasks, 5 minutes to process a document received, 10 minutes to update records on a customer and 15 and 30 minutes to make payments to suppliers, normal and high-value, respectively. Annual numbers of tasks: 60,000 documents, 24,000 customer updates, 48,000 and 2,400 payments to suppliers. Time to review an action proposed by an agent: 5, 10 or 20 minutes, with 10 as the base assumption and half as much again for high-value payments; the companion paper's time figures used 20 minutes.[21] Input documents will contain instructions to perform a malicious action in 0.1, 0.5 or 2 percent of cases. All values assumed for losses may be varied up and down by a factor of two. The most important assumed value is probably the frequency of injection, for which there is little direct information available; published measures refer to the proportion of attempts to inject instructions that are successful. All values assumed are set out in an Appendix.

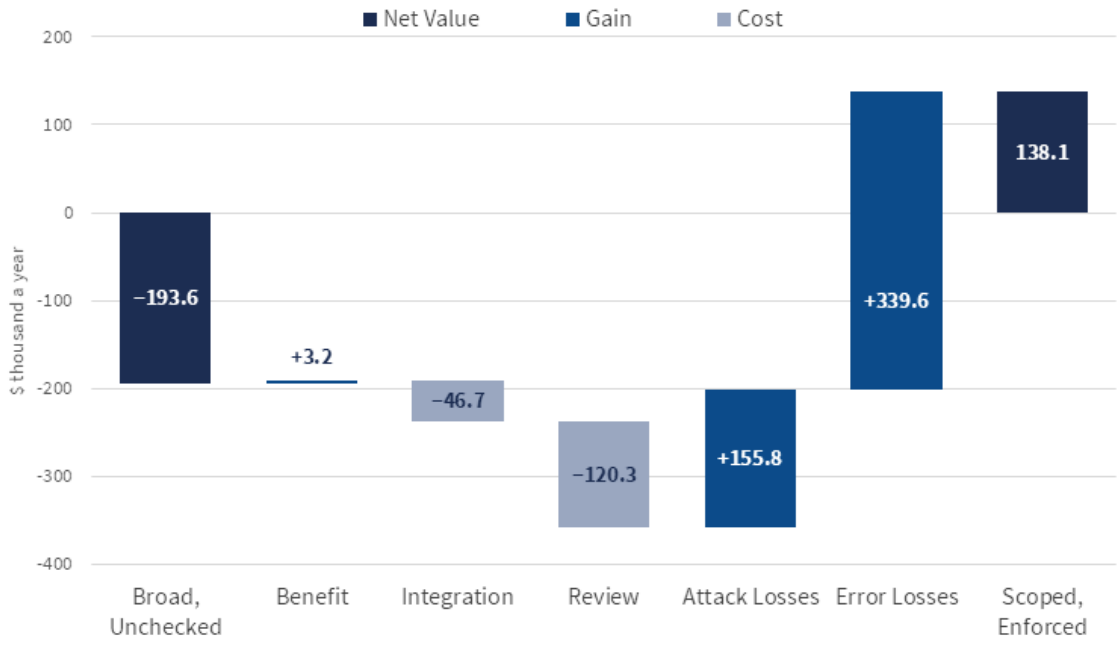
It should be noted, first, that the benefit calculated is only the reduction of time assumed to be taken if the task were performed manually. There is no allowance for errors made by staff, which would be reduced if the completion rate is increased. The best available published measurement of the effect of use of generative AI in other clerical work indicates an increase in issues successfully handled per hour of 15 percent on average for 5,172 staff responding to queries from customers, with greater effect for less experienced staff.[22] Secondly, a considerable proportion of total cost may be attributed to development and integration, for which there is even less direct information. Brynjolfsson, Rock and Syverson suggest very substantial investment in other areas, largely unobserved as intangibles, will be required to make effective use of general purpose technologies; they would depress measured productivity initially and increase it subsequently.[23] One would expect an organization already making substantial use of IT and in particular with an effective directory facility and documented policies on payments, to require less expenditure than others.

For typical supplier payments, the result with maximum authorization and no policy check is negative, at minus $193,600 per annum. The gross benefit of time saved, $390,100, is far less than the total expected loss of $529,500, of which around 65 percent, $343,400, results simply from errors in paying the inappropriate supplier and the cost of remedial action. Restricting authorization and applying policy checks gives a positive result of $138,100. The additional cost of integration and review is $46,700 plus $120,300 respectively, reducing expected loss by

$155,800 due to injection and $339,600 from error. So, on these assumptions, the policy check would be profitable in this case largely independent of any attack, depending primarily on the reduction of ordinary errors and time spent correcting them. This conclusion would be robust to substantially lower rates of injection, but would reverse for sufficiently low cost of error, including correction.

Figure 6. Avoided error losses pay for the controls more than avoided attack losses do.

Change in net value, routine supplier payments, $ thousand a year



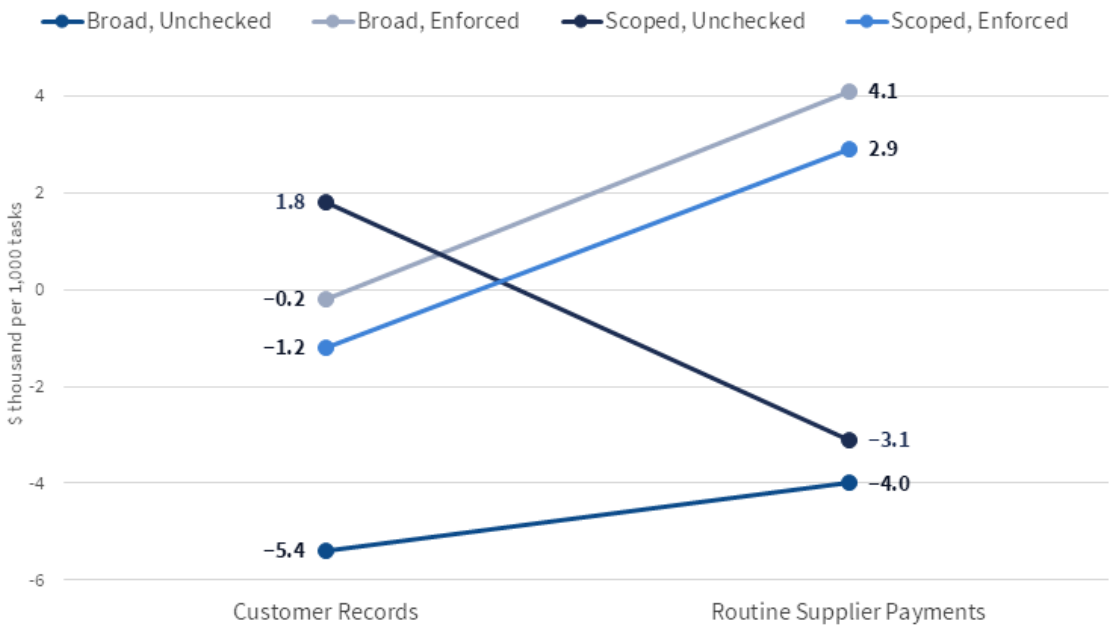
Note: Model simulation under stated assumptions, with scenario money values. 48,000 payments a year, injection prevalence 0.5% and 10 minutes per approval; compute changes by less than $100.

Source: The Economy Research Editorial (2026) [AI and Cyber-Security] The economics of delegating authority to AI: procurement, review capacity and break-even controls. The Economy Research paper; SIAI Research Editorial (2026) [AI and Cyber-Security] Agent-level versus system-level safety. SIAI Research working paper.

It can be seen that the best configuration varies according to the type of process and the nature of loss likely to occur. For updating records on customers, the main assumed loss is the inappropriate release of personal information, which is largely eliminated by restricted user IDs. So this is preferred in all 27 cases, varying rate of injection, time to review and scale of loss, giving a positive net benefit of $1,845 per 1,000 tasks at the central values. An additional policy check would increase the cost of review by $44,000 per annum, for little further reduction in expected loss. In contrast, for typical payments to suppliers, it is preferable to have maximum authorization with a policy check, giving $4,060 per 1,000, compared with negative $3,093 for restricted authorization without a check. With the improved policy, maximum authorization and a policy check are preferred in all 27 cases. It is apparent that the user ID used for payment will generally have authorization sufficient to instigate a fraudulent payment. Increasing the rate of injection from 0.1 to 2 percent reduces net benefit per 1,000 tasks with maximum authorization and no policy check from minus $1,207 to minus $21,133 for customer records and from minus $953 to minus $15,588 for payments; in contrast, the best configuration loses less than $2,600 in each case.

Figure 7. Scope alone pays best in customer records, and enforcement pays best in supplier payments.

Net value by configuration, $ thousand per 1,000 tasks



Note: Model simulation under stated assumptions, with scenario money values. Injection prevalence 0.5% and 10 minutes per approval.

Source: The Economy Research Editorial (2026) [AI and Cyber-Security] The economics of delegating authority to AI: procurement, review capacity and break-even controls. The Economy Research paper; SIAI Research Editorial (2026) [AI and Cyber-Security] Agent-level versus system-level safety. SIAI Research working paper.

None of the configurations gives a positive net benefit for processing documents received, on these central assumptions. The relatively small saving in time per document, 5 minutes, would not support a base cost of integration of $50,000 per annum and the further cost of review resulting from policy checks. There are some combinations of short review times and low rates of injection and losses for which the result would be positive. Similarly, all options give negative results for high-value payments, as the assumed annual number, 2,400, is too small to cover costs. Maximum authorization with policy check gives around $4.16 net per payment, so it would require approximately 19,600 per annum merely to cover the cost of integration. The high rate of completion of tasks in the simulation should not be taken as supporting the desirability of using an agent to perform these tasks.

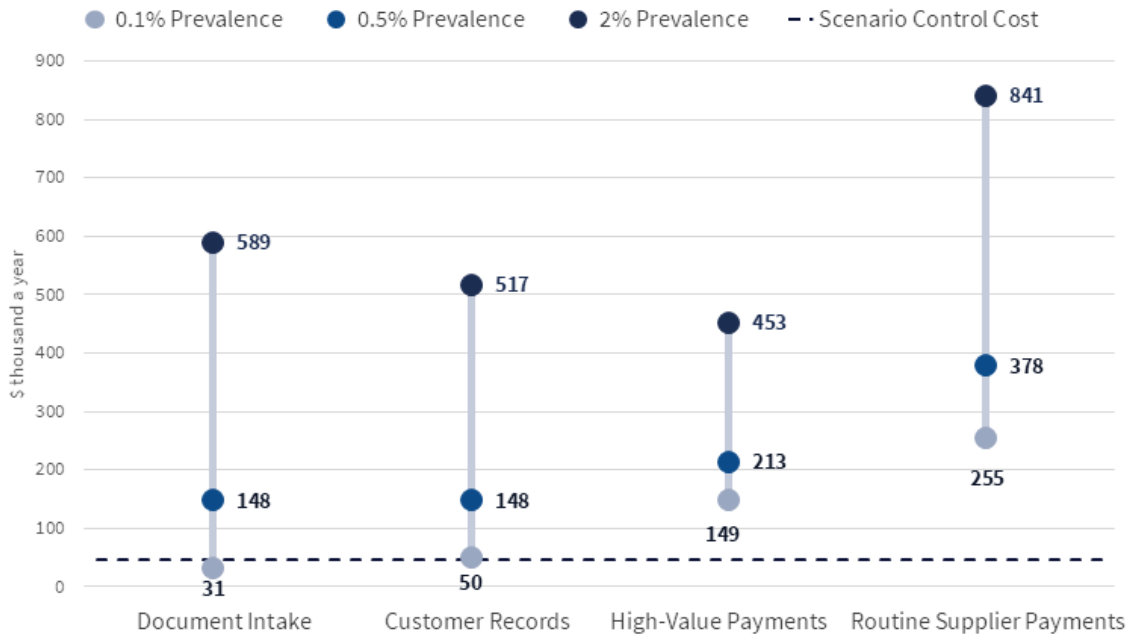
5 Where a Stronger Control Pays for Itself

As indicated, it is possible to calculate the maximum acceptable cost of additional control. This is defined as the net benefit of improved measures compared with maximum authorization and no policy check, that is, all components except the additional cost of integrating the controls. So, for restricted user ID with policy check this amounts to $148,400, $148,300, $378,400 and $212,900 per annum for documents, customer records, normal and high-value payments, respectively. On the assumptions in Chapter 4, the total cost of these measures of $46,700 per annum is well within this limit for all processes with the base rate of injection

of 0.5 percent. Only for documents received is it exceeded, with the lowest rate of injection, where break-even is $31,000. Considering all 27 combinations of assumptions, restricted authorization and policy check are preferable to no control in 74 percent of cases for documents received, 81 percent for customer records, 96 percent and 100 percent respectively for normal and high-value payments. The break-even level remains highest at every rate of injection in the case of normal payments, ranging from $255,100 to $840,800, reflecting the substantial benefit of reduced simple error.

Figure 8. The stronger configuration pays for itself everywhere except document intake at the lowest prevalence.

Break-even control budget by injection rate, $ thousand a year



Note: Model simulation under stated assumptions, with scenario money values. Gain before the cost of the controls, at 10 minutes per approval; the dashed line marks the scenario control cost of $46,700 a year.

Source: The Economy Research Editorial (2026) [AI and Cyber-Security] The economics of delegating authority to AI: procurement, review capacity and break-even controls. The Economy Research paper; SIAI Research Editorial (2026) [AI and Cyber-Security] Agent-level versus system-level safety. SIAI Research working paper.

However, almost all configurations involving some control are superior to none under these assumptions, so it is more informative to compare them with each other. It seems clear, within the limits of the processes simulated, that restriction of authorization is valuable where information may be passed externally, while policy check is valuable where payments may be made. Consistently, restriction only is always most profitable for customer records and policy only for payments. For documents, improved policy and check is best in 16 of 27 cases, restriction only in 7 and both with improved policy in 4. So there is no generally preferable option. It is notable that the most secure option, with the minimum proportion of attacks executed, is never preferable in terms of net benefit under the central assumptions, as restriction gives rise to additional costs of review and development, for little additional reduction in loss.

It is apparent that improvement of the rules applied by a policy can substantially reduce cost of review, without reducing security. If rules were improved so that 1, compared with 4, percent

of legitimate actions matched no rule, there would be little change to the frequency of successful execution of malicious instructions, but a reduction from around 8,000 to 2,100 requests for approval per annum for normal payments with maximum authorization and policy check. So the cost of review would fall from $76,500 to $20,400 per annum. Assuming the cost of developing improved rules is $23,300 per annum, net benefit would increase from $194,900 to $229,300. There will be some limit to the usefulness of improved rules, as sufficiently broad rules permitting variation in normal transactions would admit inappropriate actions, within authorized scope, which policy would not identify. Simulations have not established this point.

Time taken per review will be an important determinant of the relative benefit of maximum and restricted authorization, more so than rate of injection. So for normal payments, net benefit per 1,000 tasks with restricted authorization and policy check ranges from $4,130 for 5 minutes per approval to $372 for 20 minutes, compared with $4,860 and $2,466, respectively, without restriction of authorization. An organization able to provide additional management time for review would be able to support the higher level of security of restricted authorization. On the other hand, this implies that the high level of accuracy assumed of staff reviewing actions proposed by an agent would be required. For high-value payments, the benefit of controls exceeds $200,000 per annum, but as indicated, it would not be worthwhile to delegate the function at assumed annual numbers, so an organization would either continue manual operation or include such payments within an overall agent-supported process.

6 Conclusion

These simple calculations indicate that most of the value or otherwise of delegating tasks to an AI system will depend on decisions far removed from the selection and procurement of a model. In particular, user ID authorization and policy checks at the point of action make a substantial difference, of the order of hundreds of thousands of dollars per annum in the case considered of payments to suppliers. There would be little difference between two levels of model capability once a policy check is in place. It appears that appropriate measures will vary according to processes: restricted authorization where information may pass out of the organization, policy checks over actions making payment. In the case of normal payments, much of the benefit is from reduction in simple error. There will be tasks, typically with low benefit per transaction or low annual volume, for which no option appears attractive.

Various conclusions as to the appropriate organization are less clear-cut. Staff approving action proposed by an agent will normally have insufficient information, so that real authority remains elsewhere. The volume of approvals required will depend largely on the extent of normal transactions excluded by rules of a policy. It appears that various functions necessary for appropriate control and response to failure would normally be dispersed between several areas of an organization, each lacking full information. Most organizations would bear the direct cost of fraudulent payment not recovered from others. Introduction of verification of payee for credit transfer within the Eurozone may mitigate some of the risk from fraudulent amendment of records. A further paper considers possible incentives for others to provide improved security of AI systems generally.[24] At an individual organization level, the question arises of an appropriate function with overall responsibility. On present reasoning, it may be that no single function is so responsible.

Finally, it should be emphasized that all rates of inappropriate behavior of models are simply assumed for purposes of simulation, based on published measures of vulnerability to attack. Similarly, all other values of cost and loss and rates of injection are simply assumed. It would be unreasonable to attach any significance to absolute levels of results. However, the relative merits of different options appear robust across a wide range of assumptions. It has not been possible to establish the accuracy with which staff would approve or reject actions referred to them, in volumes of the order of 12,600 per annum; or how frequently models deployed in practice would propose inappropriate actions. Either result might substantially alter conclusions, increasing or reducing the extent to which tasks should be delegated.

References

[1, 2, 14] Association for Financial Professionals (2026) 2026 AFP Payments Fraud and Control Survey Report: Key Highlights. Rockville, MD: AFP, underwritten by Truist, April.

[3, 8, 9, 21] SIAI Research Editorial (2026a) [AI and Cyber-Security] Agent-level versus system-level safety: permission scope, action enforcement and workflow composition. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

[4] Simons, A. (2025) Announcing Microsoft Entra Agent ID: secure and manage your AI agents. Microsoft Entra Blog. Redmond, WA: Microsoft, 19 May.

[5] OWASP Foundation (2025) OWASP Top 10 for LLM Applications 2025: LLM06 Excessive Agency. OWASP GenAI Security Project.

[6] SailPoint (2025) AI Agents: The New Attack Surface. A Global Survey of Security, IT Professionals and Executives. Austin, TX: SailPoint, 28 May.

[7] Tanner, B. and Kerry, C.F. (2026) 'AI agents don't need superintelligence to cause irreversible harms', in Brookings Institution, Is the AI existential threat real – and can regulatory action prevent it? Washington, DC: Brookings Institution, 23 September.

[10] Aghion, P. and Tirole, J. (1997) 'Formal and real authority in organizations', Journal of Political Economy, 105(1), pp. 1–29.

[11] Garicano, L. (2000) 'Hierarchies and the organization of knowledge in production', Journal of Political Economy, 108(5), pp. 874–904.

[12] Committee of Sponsoring Organizations of the Treadway Commission (2013) Internal Control – Integrated Framework. Durham, NC: COSO, May.

[13] Institute of Internal Auditors (2020) The IIA's Three Lines Model: An Update of the Three Lines of Defense. Lake Mary, FL: The Institute of Internal Auditors, July.

[15] Federal Bureau of Investigation (2026) Internet Crime Report 2025. Washington, DC: FBI Internet Crime Complaint Center, April.

[16, 17] UK Finance (2026) Annual Fraud Report 2026. London: UK Finance, June.

[18] European Parliament and Council of the European Union (2024) Regulation (EU) 2024/886 of 13 March 2024 amending Regulations (EU) No 260/2012 and (EU) 2021/1230 and Directives 98/26/EC and (EU) 2015/2366 as regards instant credit transfers in euro. Official Journal of the European Union, L 2024/886, 19 March.

[19] U.S. Bureau of Labor Statistics (2026a) Occupational Outlook Handbook: Bookkeeping, Accounting, and Auditing Clerks; Accountants and Auditors. Washington, DC: U.S. Department of Labor, pay data for May 2025.

[20] U.S. Bureau of Labor Statistics (2026b) Employer Costs for Employee Compensation – June 2026. News release. Washington, DC: U.S. Department of Labor, 9 September.

[22] Brynjolfsson, E., Li, D. and Raymond, L. (2025) 'Generative AI at work', Quarterly Journal of Economics, 140(2), pp. 889–942.

[23] Brynjolfsson, E., Rock, D. and Syverson, C. (2021) 'The productivity J-curve: how intangibles complement general purpose technologies', American Economic Journal: Macroeconomics, 13(1), pp. 333–372.

[24] SIAI Research Editorial (2026b) [AI and Cyber-Security] Economic incentives behind AI safety competition: deployment rivalry, loss internalization and security effort. SIAI Research working paper, November. Swiss Institute of Artificial Intelligence.

Appendix: Scenario Values

Behavioral rates for every configuration come from a rerun of the companion paper's large replication, with the same templates, seeds and parameters, counting executed and unrecovered assets by type; its pooled results match the companion's published values. A second run changes one parameter, the share of legitimate steps that match no allow rule, from 0.04 to 0.01, to represent an improved policy. Money values are scenario values except for the published rates of pay given in Chapter 4.

Table 2. A1. Scenario Values by Workflow

Input	Documents received	Customer records	Normal supplier payments	High-value payments
Simulated task family	documents	records	payments	payments
Tasks a year	60,000	24,000	48,000	2,400
Manual minutes per task	5	10	15	30
Minutes per approval, base	10	10	10	15
Loss per disclosed asset, attack / error	$1,500 / $250	$4,000 / $500	$2,000 / $500	$2,000 / $500
Loss per unrestored record or document, attack / error	$300 / $150	$800 / $300	$500 / $250	$500 / $250
Loss per unrecovered payment to a wrong payee, attack / error	$8,000 / $800	$8,000 / $800	$8,000 / $800	$120,000 / $12,000
Loss per unrecovered payment manipulated within scope	$2,400	$2,400	$2,400	$36,000

Note: All values are assumed; losses apply to unrecovered assets.

Table 3. A2. Scenario Values Common to All Workflows

Common input	Value
Horizon	three years, up-front costs spread evenly
Total hourly cost, clerk / accountant	$34.80 / $57.47
Compute per model invocation	$0.02
Base integration	$90,000 up front, $20,000 a year
User ID restriction	$30,000 up front, $5,000 a year
Policy check	$50,000 up front, $15,000 a year
Policy improvement	$40,000 up front, $10,000 a year
Rate of injection	0.1, 0.5 or 2 percent of tasks
Minutes per approval	5, 10 or 20, scaled by 1.5 for high-value payments
Loss scale	0.5, 1 or 2 times every loss value

Note: Hourly costs are May 2025 median wages divided by the 70.0 percent wage share of compensation; other values are assumed.
Source: U.S. Bureau of Labor Statistics (2026a; 2026b).