Optimal Integration: Human, Machine, and Generative AI

Hongda Zhong, University of Texas at Dallas, CEPR, and FTG*

January 18, 2025

Abstract

I study the optimal integration of humans and technologies in multi-layered decision-making processes. Each layer can correct existing errors but may also introduce new ones. A one-dimensional quality metric – a decision-maker's error correction capability normalized by its new errors – determines the optimal rule: deploying higher-quality technologies in later stages. Interestingly, the final decision-making layer may not achieve the greatest error reduction; instead, its role hinges on minimizing new errors. Human effort varies asymmetrically across layers—early stages prioritize error correction with lower effort, while later stages emphasize avoiding new errors with higher effort. Applying the model to artificial intelligence (AI) reveals that AI's generative capabilities make it more likely to serve as the final decision-maker, reducing the need for costly human input, but underscoring the risks of AI hallucination. The theoretical framework also extends to applications including repeated delegation, automation design, loan screening, tenure review, and other multi-layer decision-making scenarios.

Key Words: Markov matrices, Multi-layer decision making, Error reduction, Automation, Delegation

JEL Codes: C44, M51

*For valuable discussions and comments, I thank Qiushi Huang, Alain Bensoussan, Jonathan Berk, Mike Burkart, Philip Bond, Will Cong, Ron Kaniel, Ivo Welch, Lucy White, Xingtan Zhang, and seminar and conference participants at UT-Dallas, CKGSB, Cavalcade Asia-Pacific conference, and Asia Finance Theory Group meeting (virtual presentation). All errors are my own.

1 Introduction

Modern production and decision-making processes increasingly depend on automation, with generative artificial intelligence (AI) technologies such as ChatGPT, autonomous vehicles, and FinTech lending experiencing rapid growth in recent years. While technologies often reduce human errors and save human effort, they can also introduce new errors into the system. This raises several key questions: How should humans and technology be optimally integrated? Who should serve as the ultimate decision-maker? How does this integration impact human effort?

To address these questions, I propose a framework based on Markov matrices to analyze the optimal integration of humans and technologies in multi-layered decision processes. Each decision-making entity—whether human or technology—takes the recommended action from the preceding layer as input and generates a (potentially random) recommendation for the next layer. The recommendation of the final layer determines the ultimate action, with payoffs potentially depending on unobservable states.

To illustrate the framework, let us consider a binary action scenario in automation, such as aviation autopilots or autonomous vehicles. Actions are broadly classified into two categories: safe or dangerous. A decision-maker (human or automated technology) can make two types of errors: altering a safe action to a dangerous action (type-1 error, with probability e_1) and failing to correct a dangerous action into a safe one (type-2 error, with probability e_2). Optimal integration seeks to maximize the probability of adopting a safe action and provides guidance for the allocation of authority between humans and technologies.

Mathematically, the effect of a decision maker is captured by a 2×2 probability transition matrix spanned by the two error probabilities $\mathcal{M} \equiv \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2 & e_2 \end{pmatrix}$. Multiplying the input action distribution by this Markov matrix yields the recommended action distribution for the next layer. The optimal integration is a particular sequence to multiply these matrices (or equivalently, applying the corresponding technologies) in order to maximize the probability of ultimately choosing the correct action.

The rule of thumb for optimal integration is to deploy higher-quality technologies in later stages. Intuitively, superior technologies act as "gatekeepers," preventing subsequent inferior technologies from altering safe actions erroneously, thereby improving outcomes.

However, defining a one-dimensional "quality" metric to rank technologies in multi-dimensional contexts (e.g. two types of errors e_1 and e_2) is not straightforward and may not even be feasible. A key contribution of this work is to establish such quality metrics for economically relevant applications, while demonstrating the absence of a universal metric in general cases.

In the binary action case, the invariant probability ($\frac{1-e_2}{1-e_2+e_1}$) associated with the transition matrix $\mathcal{M}$ serves as a one-dimensional quality metric, uniformly ranking decision technologies. This metric also has an intuitive meaning: It reflects a technology's error correction capability ($1-e_2$) normalized by the amount of new errors it introduces (e_1). Applying a technology shifts the prior probability of a safe action towards its invariant probability. Optimal integration progressively increases the likelihood of achieving a safe action throughout the decision-making process.

An intriguing observation is that the final decision-maker, equipped with the highest-quality technology (as measured by $\frac{1-e_2}{1-e_2+e_1}$ or equivalently $\frac{1-e_2}{e_1}$), may not necessarily contribute the most to error reduction. Instead, their primary role is minimizing the introduction of new errors (ensuring a low type-1 error e_1).

Next, I examine how human efforts to mitigate errors are influenced by their placement within the integration process. Effort incentives across layers reveal distinct patterns for reducing type-1 and type-2 errors. While both types of effort increase in later layers, the effort to reduce type-1 errors rises more sharply. Early agents exert lower effort and focus on reducing type-2 errors (correcting existing errors), whereas final agents exert higher effort and prioritize reducing type-1 errors (avoiding new errors).

Intuitively, effort incentives are shaped by two factors: the "*relevance*" of the error type faced by the human decision-maker and the "*consequences*" of human errors for the correctness of the final action. Errors in later stages carry greater consequences due to fewer subsequent layers available for a potential correction. However, the relevance of error types is asymmetrical. Type-1 errors, arising from correct inputs, become more relevant in later stages, where the input actions are more likely to be correct. Conversely, type-2 errors, arising from incorrect inputs, become less relevant as erroneous inputs diminish. Consequently, incentives for mitigating type-1 errors grow throughout the process, driven by both increasing relevance and consequence, while incentives for type-2 errors rise more slowly, as greater consequences outweigh reduced relevance.

The problem is significantly more complex with multiple states (denoted by T) and feasible actions (denoted by A). This complexity accommodates state-contingent errors of varying severities. For instance, in loan screening, banks make $A = 2$ decisions—accepting or rejecting applicants—who belong to $T = 2$ types: creditworthy or unworthy. Accepting unworthy borrowers can incur significant default losses, while rejecting creditworthy applicants results in milder losses from foregone interest. Generic decision technologies (e.g., human loan officers or automated lending systems) can be modeled using a probability transition matrix in the state-action space (e.g., a 4×4 matrix when $A = T = 2$).

For binary actions and multiple states ($A = 2$ and $T \geq 2$), I derive an explicit condition for determining the optimal integration of any two technologies, independent of initial action distributions. This condition incorporates expected error costs and the relative quality of the technologies across states. Surprisingly, this binary relation is not transitive when more than two technologies are considered. For example, three technologies $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ can form a circular order where $\mathcal{M}_1\mathcal{M}_2$, $\mathcal{M}_2\mathcal{M}_3$, and $\mathcal{M}_3\mathcal{M}_1$ are each optimal depending on the pairing.

This lack of transitivity indicates that no one-dimensional "quality metric" can universally determine the optimal integration of arbitrary technologies. For multiple actions ($A \geq 3$), the impossibility of a universal one-dimensional rule is further confirmed. Additionally, optimal integration may depend on initial action distributions, adding further complexity. I hope future research will provide deeper insights into these general cases.

By imposing additional structures on the decision matrices, I extend the general framework to analyze two applications: human-AI integration and multi-layer delegation.

In the first application, both humans and AI are subject to type-1 and type-2 errors. However, unlike humans, who often require costly effort to actively make decisions, AI can handle large volumes of tasks relatively inexpensively. Moreover, generative AI can sometimes produce outcomes beyond typical human capabilities.[1]

The analysis offers a simple criterion for AI to assume final decision authority. The more creative and accurate the AI, the greater its likelihood of becoming the ultimate decision-maker. However, delegating the final decision to AI does not require

[1]Notable examples include AlphaGo's victory over human players and ChatGPT's ability to write poetry surpassing most individuals.

its error correction capability to be better than human. In contrast, AI hallucinations that introduce new errors (type-1 errors) significantly hinder its suitability for final decision-making roles. Additionally, the presence of AI reduces the need for humans to make active decisions, thereby lowering effort costs.

The multi-layer delegation application builds on the classic delegation model by Aghion and Tirole (1997). A principal delegates project choices to agents who may possess superior decision-making skills but may also have preferences misaligned with the principal's, representing varying levels of loyalty. Agents can also make mistakes by incorrectly overruling their predecessor's informed decisions.

The framework produces a one-dimensional metric based on agents' skills, loyalty, and likelihood of mistakes, ranking them to form an optimal delegation sequence. A somewhat surprising insight is that when agents are error-free, ranking depends solely on loyalty, regardless of skill levels.

Additional applications discussed include autopilot design in aviation and multi-layer approval processes, such as academic promotions, student admissions, and budget reviews.

Literature Review

This paper connects with several strands of literature in statistics, decision science, and economics. The spirit of the decision problem aligns with the statistical decision-making literature, particularly "sequential analysis" (see Johnson, 1961 for a survey of the literature). Unlike standard statistical tests that rely on fixed samples, sequential analysis allows sampling decisions to be endogenous—each observation informs whether to accept or reject the null hypothesis or to continue sampling. A key development in this area is Wald's sequential probability ratio test (e.g., Wald, 1945).

In addition, the topic of combining results from multiple tests is also relevant in biostatistics (e.g., Su and Liu, 1993; Huang et al., 2011). While these problems often leverage type-1 and type-2 errors (e.g., thresholds in sequential probability ratio tests or ROC curves), they typically do not embed a Markov structure, and thus the optimal arrangement of Markov matrices is not a focus. Additionally, the role of human effort in reducing errors is largely overlooked in this literature.

Relatedly, Bayesian learning and herding models in economics (Banerjee, 1992; Bikhchandani et al., 1992; Cong and Xiao, 2024) offer useful comparison. For instance, Bayesian updating with normally distributed priors and signals generates

posterior estimates as weighted averages of priors and signals based on their respective precisions. Herding models, by contrast, assume subsequent decision-makers observe only prior actions rather than signals. My setting is closer to herding models in that each decision maker's recommendations are based on previous actions rather than signals. However, these frameworks similarly lack Markov property, and the optimal sequencing of players is not their typical focus.[2]

The use of Markov matrices in this paper links to the extensive literature on Markov processes (e.g., Stroock, 2013 as a textbook reference) and Markov decision theory, which is a staple in dynamic economic models. However, while the literature on Markov process typically analyzes the properties of given stochastic processes (e.g., periodicity, ergodicity, reversibility, convergence, or optimal stopping), this paper focuses on optimizing the sequence of multiplying different Markov matrices.

Within decision science, the Analytic Hierarchy Process (AHP) in Multiple-Criteria Decision Making (MCDM), pioneered by Saaty (1977), offers a robust method for selecting the best option among candidates with multi-dimensional attributes. My work complements AHP by focusing on determining the optimal sequence of applying different decision-making technologies. While both approaches utilize matrix formulation, they address distinct problems. AHP uses eigenvectors to prioritize attributes based on pairwise preferences, whereas my analysis uses them to optimally sequence two-dimensional Markov matrices as one of the results.

From an application perspective, this paper also contributes to the literature on effort problems involving multiple agents. Holmstrom (1982) examines moral hazard in teams but does not incorporate hierarchical structures. In addition, this paper introduces a novel separation of two types of effort—correcting mistakes and improving execution—and demonstrates how these efforts evolve differently across decision-making layers.

Two strands of economic literature address multi-layered structures: organizational hierarchies in firms (e.g., Calvo and Wellisz, 1979; Qian, 1994; Chen, 2017; Garicano, 2000) and intermediation chains in financial markets (e.g., Glode and Opp, 2016; Glode, Opp, and Zhang, 2019; He and Li, 2023; Dasgupta and Maug, 2021). This paper contributes by introducing the concept of optimally sequencing technolo-

[2]A notable exception is Bikhchandani et al. (1992), where they show that it is beneficial for "fashion leader" to have less precise information in order to defer the emergence of herding. While related to the optimal sequence, their analysis does not solve for the optimal arrangement of an arbitrary set of heterogeneous decision makers.

gies and applying it to artificial intelligence contexts.

Methodologically, this paper builds on Zhong (2023), where a special (two-dimensional) case of this framework analyzes skill and effort distribution in intermediation chains. The current framework generalizes these ideas, incorporating multiple states and actions while shifting focus to the optimal integration of decision-making layers.

The rest of the paper is organized as follows. Section 2 introduces the baseline setting for technological integration and provides several examples. Section 3 solves for the optimal integration rule in the case of $A = 2$ actions and demonstrates the absence of a one-dimensional rule in the general case ($A \geq 3$). Section 4 incorporates human effort to modify decision technologies and examines how human incentives are influenced by their position in the decision chain. Finally, two applications of the framework — human-AI integration and multi-layer delegation — are explored in Section 5, followed by a technical extension in Section 6.

2 Baseline Model Setup

2.1 Model

Consider a multi-layer decision-making process with T unobservable fundamental states ($\mathbb{T}$) and A available actions ($\mathbb{A}$). The probability distribution of fundamental states is denoted by $\mathbf{q} = (q_1, q_2, ..., q_T)$, while the payoff from action a taken in state t is represented by U_{at}. Collectively, these payoffs form the vector

$$\mathbf{U} = (U_{11}, U_{21}, ..., U_{A1}, U_{12}, U_{22}, ..., U_{A2}, ..., U_{1T}, U_{2T}..., U_{AT}) \in \mathbb{R}^{AT}. \tag{1}$$

The outcome of the decision-making process can be characterized by a distribution of actions $\mathbf{p}^{(t)} = (p_{1t}, p_{2t}, ..., p_{At})$, which may vary by state t. Coupled with the fundamental distribution $\mathbf{q}$, the resulting distribution on the AT dimensional state-action space is given by

$$\mathbf{P} = (q_1 p_{11}, q_1 p_{21}, ..., q_1 p_{A1}, q_2 p_{12}, q_2 p_{22}, ..., q_2 p_{A2}, ..., q_T p_{1T}, ..., q_T p_{AT}). \tag{2}$$

Therefore, the expected payoff is given by $\mathbf{PU}'$, where $'$ denotes the transpose.

There are $N \geq 2$ sequential decision makers—either human or technology. Each decision maker receives a recommended action $a_{i-1} \in \mathbb{A}$ from the previous layer as

input and produces a new recommendation a_i for the next layer. The output recommendation may be randomly distributed on $\mathbb{A}$ and may depend on the fundamental state. The initial action a_0 is drawn from an exogenous prior $\mathbf{P}_0$, and the final recommendation becomes the ultimate action.

Mathematically, each decision maker is represented by an $AT \times AT$ probability transition matrix $\mathcal{M}$ on the action-state space $\mathbb{A} \times \mathbb{T}$, such that the posterior distribution of the output recommendation is $\mathbf{P}_1 = \mathbf{P}_0\mathcal{M}$. It captures arbitrary ways that a decision maker may propose actions based on the underlying state (e.g. from private information about the fundamental state) and the input action from the previous layer. Although not considered in this paper, the setting also allows the possibility that a decision maker can change the fundamental state.

When N decision makers $\{\mathcal{M}_n\}$ are sequentially applied, the final posterior distribution is

$$\mathbf{P}_N = \mathbf{P}_0 \prod_{n=1}^{N} \mathcal{M}_n, \tag{3}$$

generating an expected payoff of

$$\mathbf{P}_N\mathbf{U}' = \mathbf{P}_0 \prod_{n=1}^{N} \mathcal{M}_n\mathbf{U}'. \tag{4}$$

An integration of a subset of the N decision makers is defined by a mapping $\sigma : \{1, 2, ..., N\} \to \{0, 1, 2, ..., I\}$ for some $I \leq N$ such that $\sigma^{-1}(i)$ uniquely exists for each $i = 1, 2, ..., I$. The mapping $i = \sigma(n)$ specifies the layer index of technology n in the integration, with $\sigma(n) = 0$ indicating exclusion. Hence, only $I \leq N$ technologies are used in the decision-making process.

The optimal integration $\sigma^*(\cdot)$ for a given prior distribution $\mathbf{P}_0$ and payoff vector $\mathbf{U}$ maximizes the expected payoff

$$\max_{\sigma(\cdot),I} \mathbf{P}_0 \prod_{i=1}^{I} \mathcal{M}_{\sigma^{-1}(i)}\mathbf{U}'. \tag{5}$$

I consider two problems:

1. **Optimal Integration of Technologies:** Determining the best sequence of exogenous decision matrices $\mathcal{M}_n$ in Section 3.

2. **Impact of Technology on Human Effort:** Understanding how the presence of technologies influences human effort in Section 4. This feature is modeled through endogenous matrices $\mathcal{M}(\mathbf{f})$ that depends on multi-dimensional effort input $\mathbf{f}$.

I then apply the framework to analyze human-AI integration and multi-layer delegation in Section 5.

It is worth noting that representing a decision maker by a transition matrix $\mathcal{M}$ essentially assumes Markov property: Given the most recent recommendation received by the decision maker, earlier inputs are irrelevant. This assumption is justified in scenarios where:

1. Only $N = 2$ decision makers are involved (e.g. human v.s. technology). In this case, the Markov property is naturally satisfied.[3]

2. Later decision makers cannot fully observe, comprehend, or process all preceding recommendations due to cognitive or practical constraints.

I provide several examples of $\mathcal{M}$ in different applications in the next subsection and discuss the relevance of Markov property.

2.2 Illustrative Examples

2.2.1 Automation: Autopilot and Self-Driving

In the simplest case with $T = 1$ (single fundamental state) and $A = 2$ (two possible actions), the model captures error correction in decision-making. Automation in aviation (e.g., autopilot) or autonomous driving provides a concrete example. Here, the two actions are abstractly defined as safe and disastrous. A safe action yields a normalized payoff of 1, while a disastrous action results in a payoff of 0.

Human operators (pilots or drivers) are prone to errors that can lead to disastrous outcomes. Automation helps reduce these risks but introduces its own vulnerabilities, such as sensor failures or inducing human complacency.

A decision maker (human or autopilot) can be represented by a 2×2 Markov matrix

$$\mathcal{M}_n = \begin{pmatrix} 1 - e_{1,n} & e_{1,n} \\ 1 - e_{2,n} & e_{2,n} \end{pmatrix}, \tag{6}$$

[3]From the perspective of the second decision maker, the only predecessor is the first one.

where $e_{1,n}$ denotes the probability of a type-1 error – changing a safe action to a disastrous one and $e_{2,n}$ denotes the probability of a type-2 error – failing to correct a disastrous action. Formally,

$$P(a_n \text{ is disastrous} | a_{n-1} \text{ is safe}) = e_{1,n}$$

and

$$P(a_n \text{ is disastrous} | a_{n-1} \text{ is disastrous}) = e_{2,n}$$

The model provides insights into optimal integration of automation, its impact on human effort, and the allocation of decision-making authority between humans and machines.

The Markov property is likely satisfied if each decision maker only observes the immediate predecessor's recommendations (e.g., a human pilot might only see the autopilot's actions without access to the flight computer's inputs from various sensors).

2.2.2 Loan Screening

This example considers $T > 1$ payoff relevant states and $A = 2$ actions. In loan screening, the fundamental states are the borrower types: Good (G) and Bad (B) for simplicity. The two possible actions are to Accept (A) or Reject (R) the borrower. The proportion of good borrowers is q_G, and denote by $q_B = 1 - q_G$.

There are $A \times T = 4$ outcomes: accepting good, rejecting good, rejecting bad, accepting bad borrowers, denoted by $\{GA, GR, BR, BA\}$ respectively. Accepting a good borrower generates an interest income of $r > 0$, while accepting a bad borrower results in a loss of principal of $-L < 0$. Rejecting a borrower yields a payoff of 0 regardless of type. Therefore, the payoff vector is $\mathbf{U} = (r, 0, 0, -L)$,

A loan screening technology (human loan officer or automated screening system) is represented by a Markov matrix on the four outcomes:

$$\mathcal{M}_n = \begin{pmatrix} 1 - e_{1,n}^{(G)} & e_{1,n}^{(G)} & 0 & 0 \\ 1 - e_{2,n}^{(G)} & e_{2,n}^{(G)} & 0 & 0 \\ 0 & 0 & 1 - e_{1,n}^{(B)} & e_{1,n}^{(B)} \\ 0 & 0 & 1 - e_{2,n}^{(B)} & e_{2,n}^{(B)} \end{pmatrix}, \tag{7}$$

where $e_{1,n}^{(\theta)}$ and $e_{2,n}^{(\theta)}$ ($\theta = G$ or B) are the type-1 and type-2 errors of technology n

when facing a type-θ borrower.

To elaborate, for a good borrower, the "correct" action—one that yields the higher payoff in that state—is to accept. Hence, a type-1 error in this case is to switch from accepting (GA) to rejecting (GR) and a type-2 error is to maintain the GR state.

In contrast, for a bad borrower, the correct action is to reject. Therefore, a type-1 error in this case is to switch from rejecting (BR) to accepting (BA) and a type-2 error is to maintain the BA state.

It is reasonable to assume that a loan screening technology can only change lending decisions, but not the borrower type. Hence, the decision matrix $\mathcal{M}_n$ only contains $T = 2$ diagonal blocks, each of size 2×2 ($A \times A$), indicating no transition between good and bad type borrowers.

More generally, a decision technology that can alter the proposed action, but not the fundamental states is represented by a block diagonal matrix

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}^{(1)} & 0 & 0 & 0 \\ 0 & \mathcal{M}^{(2)} & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \mathcal{M}^{(T)} \end{pmatrix}. \tag{8}$$

The state-dependent blocks $\mathcal{M}^{(t)}|_{A\times A}$ allow for the possibility that decision makers perform differently in different states.

The theoretical result for this case (in Subsection 3.2) mostly focuses on the case with only $N = 2$ layers, and the Markov property is automatically satisfied.

2.2.3 Multi-Layer Approval Process

The model also applies to multi-layered approval processes, such as academic promotion, budget reviews, and student admissions. In these cases, multiple layers of decision-making occur sequentially, with recommendations being passed between layers (e.g., external evaluators, committees, provosts, and presidents).

While late-stage decision makers in these cases have access to earlier information, the Markov property may still hold if each decision maker mostly reacts to the immediate prior recommendations. For instance, in academic promotion, higher-level officials (e.g. provosts or presidents) often lack the time or expertise to evaluate all prior recommendations (e.g. all reference letters) and instead depend on summaries

or condensed evaluations from their immediate predecessors.

Effort-based matrices $\mathcal{M}(\mathbf{f})$ discussed in Section 4, provide further insights into how each layer should prioritize specific aspects of the decision-making process.

3 Optimal Integration of Technologies

I first provide a succinct integration rule in Subsection 3.1 for the simplest case with binary actions ($A = 2$) and a single underlying state ($T = 1$), motivated by the application of automative technologies in Subsection 2.2.1. The case with multiple states ($A = 2$ and $T > 1$) is studied in Subsection 3.2. To complete the analysis, Subsection 3.3 provides insights on the complexity of the general problem with many actions ($A > 2$).

3.1 Binary Actions ($A = 2$) and a Single State ($T = 1$)

I start by analyzing the simplest case with binary actions ($A = 2$) and a single state ($T = 1$), using the explicit form of matrix $\mathcal{M}_n$ in (6) and the language of type-1 and type-2 errors, $e_{1,n}$ and $e_{2,n}$. This language is without loss of generality in the case of binary actions (even with multiple states as in Section 3.2), since the "correct" action in each state can always be labeled as the one generating the higher payoff. Additionally, the payoff vector is normalized to $\mathbf{U} = (1, 0)$ without loss of generality.

Throughout the paper, I assume type-2 error is higher than type-1 error

$$e_2 > e_1. \tag{9}$$

Correcting a mistake is generally more difficult than maintaining a correct action, a scenario that is arguably more relevant in most applications.[4] The necessity of this assumption will transpire shortly following condition (15), and the consequence of its violation will be studied more carefully in Subsection 6.

I first characterize the impact of applying a technology $\mathcal{M}_i$ on the success proba-

[4]For example, in aviation (Subsection 2.2.1), few things can go wrong during routine flight when the input action is correct (lower type-1 error). In contrast, recovering from dangerous actions (e.g. in an emergency when autopilot fails) is considerably more challenging (higher type-2 error).

bility. Define the invariant probability p_i^* associated with $\mathcal{M}_i$ as

$$\begin{pmatrix} p_i^* & 1-p_i^* \end{pmatrix} = \begin{pmatrix} p_i^* & 1-p_i^* \end{pmatrix} \mathcal{M}_i. \tag{10}$$

Equivalently,

$$p_i^* = (1-e_{1,i})\, p_i^* + (1-e_{2,i})\,(1-p_i^*)\,. \tag{11}$$

which simplifies to:

$$p_i^* = \frac{1-e_{2,i}}{1-e_{2,i}+e_{1,i}}. \tag{12}$$

Intuitively, this represents the probability of a correct action that remains constant after applying technology $\mathcal{M}_i$.

The evolution of probability (3) can be recursively expressed as

$$\mathbf{P}_i \equiv (p_i, 1-p_i) = \mathbf{P}_{i-1}\mathcal{M}_i,$$

or explicitly:

$$p_i = (1-e_{1,i})\, p_{i-1} + (1-e_{2,i})\,(1-p_{i-1})\,, \tag{13}$$

where p_i denotes the probability of the correct action being proposed by the i-th decision maker.

The difference between (11) and (13) is

$$p_i^* - p_i = (e_{2,i}-e_{1,i})\,(p_i^* - p_{i-1})\,, \tag{14}$$

or equivalently,

$$p_i = p_i^* - (e_{2,i}-e_{1,i})\,(p_i^* - p_{i-1})\,. \tag{15}$$

Condition (15) yields an intuitive and important observation: Technology $\mathcal{M}_i$ shifts the prior probability p_{i-1} towards its invariant probability p_i^*, and the difference between these two probabilities diminishes by a factor of $e_{2,i}-e_{1,i}$. Therefore, the posterior probability p_i lies between the prior probability p_{i-1} and the invariant probability p_i^*.

It is worth noting that this observation relies on assumption (9). If the opposite assumption holds ($e_{2,i} \leq e_{1,i}$), the posterior oscillates around p_i^*: the prior p_{i-1} and the posterior p_i lie on opposite sides of p_i^*. Intuitively, a lower prior p_{i-1} increases the posterior success p_i because the technology better handles mistakes than maintaining

correctness.

I now characterize optimal integration strategy.

Proposition 1 ***[Ranking Two Technologies]*** *Suppose there are two available technologies $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $p_0 \leq p_1^* \leq p_2^*$. Then applying $\mathcal{M}_1$ first is weakly better than applying $\mathcal{M}_2$ first. Formally,*

$$\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \mathcal{M}_1\mathcal{M}_2\mathbf{U}' \geq \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \mathcal{M}_2\mathcal{M}_1\mathbf{U}', \tag{16}$$

where equality holds if and only if $p_1^* = p_2^*$.

Intuitively, to achieve a superior outcome, higher "quality" technology should be positioned later in the sequence as the "gatekeeper." Placing a good technology early is suboptimal since subsequent worse technologies may introduce more errors, compromising the correct actions made by the initial technology. However, defining "quality" is challenging due to multidimensional error profiles: one technology may create higher type-1 errors while the other generates higher type-2 errors. Proposition 1 reduces the dimensionality of this problem, showing that the invariant probability is the key metric to rank different technologies.

It is also useful to observe that ranking technologies by their invariant probability p^* is equivalent to ranking them by $\frac{1-e_2}{e_1}$, which carries clear intuition. This ratio measures a technology's capability to correct errors (troubleshooting) relative to the amount of new errors it produces.

Thanks to this one-dimensional quality metric—invariant probability p^* or equivalently $\frac{1-e_2}{e_1}$ – one can rank any number of technologies when $N > 2$. Therefore, the intuition from Proposition 1 generalizes to an optimal integration rule as follows.

Proposition 2 ***[Optimal Integration]*** *Suppose $T = 1$ and $A = 2$, and there are N technologies, each characterized by its transition matrix $\mathcal{M}_n$, $n = 1, 2, ..., N$. The invariant probability associated with $\mathcal{M}_n$ is p_n^*. The optimal integration σ^* selects only those technologies whose invariant probability exceeds the prior probability p_0; that is*

$$\sigma^*(n) = 0 \text{ if } p_n^* < p_0.$$

The remaining technologies are integrated in ascending order of invariant probabilities, meaning

$$\sigma^*(n_1) < \sigma^*(n_2)$$

if $p^*_{n_1} < p^*_{n_2}$.

To implement the optimal integration rule in Proposition 2, one simply needs to calculate the invariant probabilities p^*_n for each of the N technologies, eliminates those with $p^*_n < p_0$ because they worsen the posterior probability, and then applies the remaining ones sequentially from lowest to highest invariant probabilities.

Since each posterior probability p_i lies between the prior probability p_{i-1} and the invariant probability p^*_i, the increasing sequence of p^*_i under optimal integration implies that the sequence of p_i is also increasing. This feature is particularly relevant when considering human effort in decision-making processes.

Corollary 1 *Under the optimal integration of technologies, the probability* p_i *increases with* i.

It is important to note that having the highest-quality technology as the final decision maker does not necessarily mean it contributes the most significant improvement to the outcome. Consider the following extreme example:

The prior correct probability is $p_0 = 0$. Two technologies are given by

$$\mathcal{M}_1 = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{pmatrix} \text{ and } \mathcal{M}_2 = \begin{pmatrix} 1 & 0 \\ 0.1 & 0.9 \end{pmatrix}.$$

It is easy to calculate the corresponding invariant probabilities:

$$p^*_1 = \frac{1-0.1}{1-0.1+0.1} = 0.9 \text{ and } p^*_2 = \frac{1-0.9}{1-0.9+0} = 1.$$

The optimal integration is therefore $\mathcal{M}_1\mathcal{M}_2$. However, technology $\mathcal{M}_1$ corrects 90% mistakes whereas technology $\mathcal{M}_2$ only corrects 10% of the remaining mistakes. The key for $\mathcal{M}_2$ to be the final decision maker is its low likelihood of introducing new mistakes (specifically, 0% type-1 error), even though its contribution to error reduction is smaller than $\mathcal{M}_1$.

I conclude this subsection with a relevant observation in aviation. Type-1 error in the model reflects execution quality (the ability to maintain safe actions) whereas type-2 error measures troubleshooting capabilities (the ability to correct dangerous actions). Machines excel in execution ($e_1 \approx 0$), but struggle with troubleshooting

($e_2 \approx 100\%$), resulting in an indeterminant invariant probability ($\frac{0}{0}$). The optimal placement of these technologies is therefore highly sensitive to the specific combination of the two errors, explaining varied approaches towards automation among companies. In the design of autopilot, Airbus pioneered the "fly-by-wire" technology in commercial aviation, effectively giving the ultimate decision to the machine (flight computer). If pilot actions are deemed dangerous by the computer, the plane overrides those actions. In contrast, Boeing adhered to an alternative design philosophy for many years, maintaining that humans should retain ultimate control (Kornecki and Hall 2004).

3.2 Binary Actions ($A = 2$) and Multiple States ($T \geq 2$)

The simplest case with $A = 2$ actions and $T = 1$ type, previously analyzed, essentially assumes all errors are equally costly. In practice, the same action may be correct in some states but wrong in others. Additionally, the cost of errors can vary depending on the underlying states (recall the loan screening application in Section 2.2.2). These features are captured by considering multiple payoff-relevant states $T \geq 2$ and binary actions $A = 2$. Binary actions allow for the interpretation of "error," as one can always label the action generating a higher payoff as the correct action.

This problem becomes significantly more complex. First, I derive the optimal integration rule for $N = 2$ technologies, analogous to Proposition 1. Next, I show that a result similar to Proposition 2 does not exist: No one-dimensional metric based only on individual technologies can generally determine the optimal integration sequence for $N \geq 3$ arbitrary technologies.

Consider a generic technology in (8) for $A = 2$ actions and an arbitrary states $T \geq 2$. Denote by (t) in the superscripts ($t = 1, 2, ..., T$) the state-dependent probability distribution $\mathbf{p}^{(t)}|_{1\times 2}$, transition matrix $\mathcal{M}^{(t)}|_{2\times 2}$, and payoff vector $\mathbf{U}^{(t)}|_{1\times 2}$. Recall from (2) that q_t denotes the probability of fundamental state t, which remains constant throughout the decision-making process. Using the block diagonal form of $\mathcal{M}$, the payoff (4) can be explicitly rewritten as:

$$\mathbf{P}_N \mathbf{U}' = \sum_{t=1}^{T} q_t \mathbf{p}_0^{(t)} \prod_{n=1}^{N} \mathcal{M}_n^{(t)} \mathbf{U}^{(t)\prime}.$$

Denote by $(\overline{u}^{(t)}, \underline{u}^{(t)}) \equiv \mathbf{U}^{(t)}$ the payoffs associated with the two actions, and

without loss of generality, assume $\overline{u}^{(t)} \geq \underline{u}^{(t)}$. Hence,

$$\mathbf{P}_N \mathbf{U}' = \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) q_t \mathbf{p}_0^{(t)} \prod_{n=1}^{N} \mathcal{M}_n^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + q_t \mathbf{p}_0^{(t)} \begin{pmatrix} \underline{u}^{(t)} \\ \underline{u}^{(t)} \end{pmatrix}. \tag{17}$$

Intuitively, each fundamental state is associated with a minimum payoff of $\underline{u}^{(t)}$, reflected by the constant second term in (17) irrespective of different integration sequence. Technologies $(\mathcal{M}_n)$ act on the payoff difference between the two actions $\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)$. Factoring this out, the payoff vector can be normalized to $(1, 0)$, yielding the first term in (17).

Ignoring the constant, the optimal integration problem becomes:

$$\max_{\sigma(\cdot), I} \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) q_t \cdot \mathbf{p}_0^{(t)} \prod_{i=1}^{I} \mathcal{M}_{\sigma^{-1}(i)}^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{18}$$

I now derive the optimal integration rule for $N = 2$ technologies. For each fundamental state t, denote by $e_{i,n}^{(t)}$ the type-i error of technology $n = 1, 2$, represented by the diagonal blocks:

$$\mathcal{M}_n^{(t)} = \begin{pmatrix} 1 - e_{1,n}^{(t)} & e_{1,n}^{(t)} \\ 1 - e_{2,n}^{(t)} & e_{2,n}^{(t)} \end{pmatrix}.$$

I calculate the payoff difference between the two orders of applying the technologies:

$$\mathbf{P}_0 \mathcal{M}_1 \mathcal{M}_2 \mathbf{U}' - \mathbf{P}_0 \mathcal{M}_2 \mathcal{M}_1 \mathbf{U}' = \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) q_t \det \begin{pmatrix} e_{1,1}^{(t)} & e_{1,2}^{(t)} \\ 1 - e_{2,1}^{(t)} & 1 - e_{2,2}^{(t)} \end{pmatrix}, \tag{19}$$

where det denotes the determinant of a matrix. The detailed calculation is relegated to the proof of the next result in the Appendix. An important feature of (19) is that the difference is independent of the initial distribution of actions $\mathbf{p}_0^{(t)}$, but only depends on the distribution of fundamental states q_t. This property allows one to rank any two technologies independent of the initial action distribution $\mathbf{p}_0^{(t)}$. Formally, we have the following result.

Proposition 3 *Suppose $A = 2$ and $N = 2$. For any $T \geq 1$, a given distribution of fundamental states q_t, and the payoff vector $\mathbf{U}$, it is optimal to apply technology $\mathcal{M}_2$*

after $\mathcal{M}_1$ *if and only if*

$$\sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) q_t \det \begin{pmatrix} e_{1,1}^{(t)} & e_{1,2}^{(t)} \\ 1 - e_{2,1}^{(t)} & 1 - e_{2,2}^{(t)} \end{pmatrix} \geq 0. \tag{20}$$

This result nests Proposition 1 (i.e., $p_2^* \geq p_1^*$) as a special case with $T = 1$, where condition (20) simplifies to:

$$\det \begin{pmatrix} e_{1,1} & e_{1,2} \\ 1 - e_{2,1} & 1 - e_{2,2} \end{pmatrix} \geq 0, \text{ or equivalently, } \frac{e_{1,1}}{1 - e_{2,1}} \geq \frac{e_{1,2}}{1 - e_{2,2}}.$$

Condition (20) is also intuitive. When comparing two technologies, their relative quality depends on the state and is measured by the determinant of their error correction capabilities ($1 - e_{2,n}^{(t)}$) and chances of making new mistakes ($e_{1,n}^{(t)}$). The overall comparison further weights the cost of making a mistake in that state $\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)$ and the likelihood of each state q_t.

When applied to the loan screening application in Section 2.2.2, condition (20) can be explicitly written as

$$rq_G \det \begin{pmatrix} e_{1,1}^{(G)} & e_{1,2}^{(G)} \\ 1 - e_{2,1}^{(G)} & 1 - e_{2,2}^{(G)} \end{pmatrix} + Lq_B \det \begin{pmatrix} e_{1,1}^{(B)} & e_{1,2}^{(B)} \\ 1 - e_{2,1}^{(B)} & 1 - e_{2,2}^{(B)} \end{pmatrix} \geq 0.$$

This condition highlights three important considerations when comparing different lending technologies (e.g. loan officers v.s. FinTech algorithms): borrower composition (q_G and q_B), the costs of errors (r or L), and the relative performance of the technologies for each borrower type measured by the determinants.

Lack of Transitivity when $N \geq 3$

Intriguingly, unlike the case of $T = 1$, the binary relation given by Proposition 3 does not possess transitivity. Formally, denote by $\mathcal{M}_2 \succ \mathcal{M}_1$ if condition (20) strictly holds, meaning that it is strictly better to apply $\mathcal{M}_2$ after $\mathcal{M}_1$ when only these two technologies are available. Then, it is possible to have circular ranking, i.e., $\mathcal{M}_3 \succ \mathcal{M}_2$, $\mathcal{M}_2 \succ \mathcal{M}_1$, and $\mathcal{M}_1 \succ \mathcal{M}_3$. The lack of transitivity also implies that there is no one-dimensional metric that can rank all technologies, forming a general integration rule like Proposition 2.

For an illustrative counterexample, consider $\mathbf{U} = (3, 2, 1, 0)$, $\mathbf{P}_0 = (0, 0.5, 0, 0.5)$, and the following three technologies

$$\mathcal{M}_1 = \begin{pmatrix} 0.75 & 0.25 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0.3 & 0.7 \end{pmatrix}, \mathcal{M}_2 = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}, \mathcal{M}_3 = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0.1 & 0.9 \end{pmatrix}.$$

One can verify that the following three relations hold:

$$\begin{aligned} \mathbf{P}_0\mathcal{M}_1\mathcal{M}_2\mathbf{U}' = 1.415 > 1.41 = \mathbf{P}_0\mathcal{M}_2\mathcal{M}_1\mathbf{U}', \\ \mathbf{P}_0\mathcal{M}_2\mathcal{M}_3\mathbf{U}' = 1.35 > 1.345 = \mathbf{P}_0\mathcal{M}_3\mathcal{M}_2\mathbf{U}', \\ \mathbf{P}_0\mathcal{M}_3\mathcal{M}_1\mathbf{U}' = 1.3875 > 1.38 = \mathbf{P}_0\mathcal{M}_1\mathcal{M}_3\mathbf{U}'. \end{aligned} \tag{21}$$

The optimal integration with all three technologies is

$$\mathbf{P}_0\mathcal{M}_3\mathcal{M}_1\mathcal{M}_2\mathbf{U}' = 1.456,$$

which dominates all other permutations of any subsets. However, condition (21) implies that in the absence of $\mathcal{M}_1$, the optimal way to integrate the remaining two technologies is to apply $\mathcal{M}_2$ first. Therefore, introducing a new technology ($\mathcal{M}_1$) can significantly disrupt how existing technologies ($\mathcal{M}_2$ and $\mathcal{M}_3$) are integrated.

Intuitively, the introduction of a new technology can alter the relative importance of each state, thereby changing the relative quality of the existing technologies. Consequently, the optimal integration may require a non-monotonic reordering of existing technologies.

This phenomenon does not arise when there is no fundamental uncertainty ($T = 1$), as the invariant probability p^* serves as a one-dimensional quality metric that ranks all technologies universally.

3.3 More than Two Actions: $A > 2$

When $A \geq 3$, the notion for "correct" or "wrong" becomes less clear, and the problem becomes even more complex. The optimal integration may not only depend on the probability of the fundamental states q_t, but also the initial distribution on actions

$\mathbf{p}_0^{(t)}$.

To illustrate this complication, consider a simple numerical example with $A = 3$ actions and $T = 1$ fundamental state. Let $\mathbf{U} = (2, 1, 0)$. The two technologies under consideration are:

$$\mathcal{M}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathcal{M}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0.5 & 0.5 & 0 \end{pmatrix}.$$

Compare two initial action distributions $\mathbf{P}_0 = (0, 1, 0)$ and $\hat{\mathbf{P}}_0 = (0, 0, 1)$. One can verify the following relations:

$$\mathbf{P}_0\mathcal{M}_1\mathcal{M}_2\mathbf{U}' = 1.75 > 1 = \mathbf{P}_0\mathcal{M}_2\mathcal{M}_1\mathbf{U}',$$

and

$$\hat{\mathbf{P}}_0\mathcal{M}_1\mathcal{M}_2\mathbf{U}' = 1 < 1.5 = \hat{\mathbf{P}}_0\mathcal{M}_2\mathcal{M}_1\mathbf{U}'.$$

Thus, the optimal integration sequence depends on the initial distribution of actions, even in the absence of fundamental uncertainty.

Intuitively, with $A \geq 3$ actions, the concept of a "correct" or "wrong" action becomes ambiguous. An action a_2 might yield a higher payoff than a_3, but perform poorly when addressing a_1. A technology might excel at improving a_2 to a_1, but bad at handling a_3. As a result, the choice of which technology to apply first can depend on the initial action distribution. The following result summarizes the discussion formally.

Proposition 4 *Except in the case where $A = 2$ and $T = 1$, no one-dimensional metric exists that can rank arbitrary technologies based solely on the characteristics of individual technologies.*

I conclude this section by explaining the technical challenge of characterizing a result analogous to (20) when $A \geq 3$ actions are available. One might conjecture that by decomposing the action space into pairwise actions, it would be feasible to transform the problem into one involving multiple $(\frac{A(A-1)}{2})$ pairs of actions, thus potentially yielding a result similar to (20). However, the issue with this approach is that the decision technologies, when limited to a pair of actions, no longer adhere to

the Markov property. Consequently, I have yet to discover an elegant criterion for the general case with $A \geq 3$. As a result, the applications discussed in Section 5, which involve multiple actions, require additional structures for the decision matrices.

4 Impact of Technology on Human Effort

So far, the analysis has focused on the optimal integration of exogenous technologies. In contrast, humans can exert costly effort to modify or improve these technologies. This section explores how technological integration influences human effort incentives. Subsection 4.1 analyzes the effort incentive of a single human decision-maker and Subsection 4.2 extends the intuition to study the case with multiple human layers and effort specialization.

To provide a tangible interpretation of effort and errors, I will focus on the binary action case discussed in Subsection 3.1.[5] An interesting result is that the incentive for human effort varies depending on their position in the integration sequence and exhibits asymmetry between type-1 and type-2 errors.

4.1 Effort Incentives of a Single Player

To formally introduce human effort, suppose that there is a strategic human decision maker among the N layers, henceforth, the "player". This player operates in layer $j \in \{1, 2, ..., N\}$ and creates type-i error $h_i(f_{i,j})$ which depend on effort input $f_{i,j}$, where $i = 1, 2$. The notation h_i highlights endogenous errors by the human player, contrasting with the exogenous errors probabilities e_i of machines. Hence, similar to (6), the player's decision technology is given by

$$\mathcal{M}_j = \begin{pmatrix} 1 - h_1(f_{1,j}) & h_1(f_{1,j}) \\ 1 - h_2(f_{2,j}) & h_2(f_{2,j}) \end{pmatrix}. \tag{22}$$

If the correct action is chosen, the player receives a positive utility normalized to 1. Otherwise, there is no utility. The cost of effort is denoted by $c_i(f_{i,j})$ $(i = 1, 2)$. Our focus in this subsection is the player's optimal effort levels, $f_{1,j}$ and $f_{2,j}$, that

[5]When $A = 2$, the action that generates a higher payoff can always be labeled as the correct one. When multiple actions ($A > 2$) are possible, the mathematical expressions are similar, but the interpretation becomes less intuitive. Effort here can be thought of as replacing an existing technology with a new one. For further discussion, see the effort analysis in Section 5.1.

maximize the expected payoff net of effort costs:

$$\max_{f_{1,j},f_{2,j}} p_N - c_1(f_{1,j}) - c_2(f_{2,j}).$$

The remaining $N-1$ technologies $\mathcal{M}_n$ ($n \neq j$) are given exogenously and optimally arranged in the ascending order of their invariant probabilities p_n^*, as specified by Proposition 2. The case of N human players is analyzed in Section 4.2. I impose standard assumptions to guarantee a solution:

- Error functions: $h_i(f_{i,j})$ are decreasing and convex in $f_{i,j}$, reflecting diminishing marginal benefit of effort.
- Costs functions: $c_i(f_{i,j})$ are increasing and convex in $f_{i,j}$, reflecting increasing marginal cost of effort.

The first-order condition for $f_{1,j}$ can be decomposed as follows:

$$\begin{aligned} c_1'(f_{1,j}^*) &= \frac{\partial \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^N \mathcal{M}_i \mathbf{U}'}{\partial f_{1,j}} \\ &= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} h_1'(f_{1,j}) \prod_{i=j+1}^N \mathcal{M}_i \mathbf{U}' \\ &= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \mathbf{U}' \cdot \begin{pmatrix} -1 & 1 \end{pmatrix} h_1'(f_{1,j}) \prod_{i=j+1}^N \mathcal{M}_i \mathbf{U}' \\ &= -h_1'(f_{1,j}^*) \cdot \underbrace{\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \mathbf{U}'}_{\text{relevance of type-1 error}} \cdot \underbrace{\prod_{i=j+1}^{N} (e_{2,i} - e_{1,i})}_{\text{consequence of type-1 error}}, \end{aligned} \tag{23}$$

where $\mathbf{U} = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

Two intuitive channels determine the marginal benefit of effort. The first channel is the "relevance" of type-1 error in layer j, captured by the term $\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \mathbf{U}'$ in (23). It represents the probability of a *correct* action is proposed to the human player, the case where type-1 error is relevant. The second channel $\prod_{i=j+1}^N (e_{2,i} - e_{1,i})$ is the "consequence" of an error in layer j, reflecting its impact on the correctness of the final action.

Similarly, the first-order condition for $f_{2,j}$ can be decomposed into the relevance

(when the input action is wrong) and consequence components:

$$
\begin{aligned}
c_2'(f_{1,j}^*) &= \frac{\partial \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{N} \mathcal{M}_i \mathbf{U}'}{\partial f_{2,j}} \\
&= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix} h_2'(f_{2,j}) \prod_{i=j+1}^{N} \mathcal{M}_i \mathbf{U}' \\
&= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \end{pmatrix} h_2'(f_{2,j}) \prod_{i=j+1}^{N} \mathcal{M}_i \mathbf{U}' \qquad (24) \\
&= -h_2'(f_{2,j}^*) \underbrace{\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\text{relevance of type-2 error}} \underbrace{\prod_{i=j+1}^{N} (e_{2,i} - e_{1,i})}_{\text{consequence of type-2 error}}
\end{aligned}
$$

Comparing (23) and (24), the consequence term is identical for both error types, but the relevance terms differ due to their dependence on the probabilities of correct and incorrect input actions.

For convenience, define the *effort incentive* for reducing type-i errors as the ratio:

$$\text{effort incentive}_i \equiv \frac{c_i'(f_{i,j}^*)}{-h_i'(f_{i,j}^*)}, \qquad (25)$$

which represents the marginal cost per unit of error reduction. This measure captures the player's equilibrium effort incentive.[6]

The decompositions (23) and (24) of effort incentives into relevance and consequence components provide a clear framework for understanding the asymmetric behavior of effort incentives for type-1 and type-2 errors.

Proposition 5 ***[Effort Incentive]*** *Suppose a human player operates in layer j of an N-layer decision making process, and the other $N-1$ technologies are optimally integrated according to Proposition 2, then the player's optimal effort levels $f_{1,j}^*$ and $f_{2,j}^*$ are increasing in j.*

Furthermore, the equilibrium effort incentive for $f_{1,j}^*$ increases more quickly than

[6]For classic linear-quadratic benefit-cost functions $c(f) = \frac{1}{2}f^2$ and $h(f) = 1 - f$ ($f \in [0,1]$), this ratio simplifies to the exact equilibrium effort choice $\frac{c'(f^*)}{-h'(f^*)} = f^*$. More generally, under natural convexity and monotonicity assumptions, expression (25) is increasing in f^*.

that for $f^*_{2,j}$. Mathematically,

$$\frac{-c_1'(f^*_{1,j+1})/h_1'(f^*_{1,j+1})}{-c_1'(f^*_{1,j})/h_1'(f^*_{1,j})} > \frac{-c_2'(f^*_{2,j+1})/h_2'(f^*_{2,j+1})}{-c_2'(f^*_{2,j})/h_2'(f^*_{2,j})} > 1. \tag{26}$$

Intuitively, for type-1 errors, effort increases later in the chain because both the relevance and the consequence of type-1 errors increase. On the one hand, type-1 errors are more relevant because the proposed action is more likely to be correct (Corollary 1). On the other hand, the consequence of an error is more severe in later layers due to fewer subsequent layers that can potentially make corrections.

The intuition is more intricate for type-2 errors. While the consequence of type-2 errors increases (by a factor of $\frac{1}{(e_{2,j}-e_{1,j})}$), their relevance decreases (by a factor of $\frac{1-p_{j-1}}{1-p_j}$), because inputs are less likely to be wrong in later layers. However, as demonstrated in the appendix, the increased consequence dominates the decreased relevance, resulting in an overall increase in effort incentive for $f^*_{2,j}$ with j albeit more slowly than $f^*_{1,j}$.

Impact of Technology on Human Effort in Practice

A common concern associated with reliance on automation, frequently highlighted in news reports, is that human operators (pilots or drivers) may lose focus or even fall asleep once the autopilot or auto-drive system takes control. Notable incidents reinforcing this issue include Northwest Airlines Flight 188 in 2009 and Batik Air Flight 6723 in 2024.[7] Such occurrences are even more prevalent in the context of autonomous driving, with multiple instances reported of Tesla drivers falling asleep at the wheel.[8] While this negligence is both dangerous and illegal, it is a natural consequence predicted by Proposition 5: Technologies capable of making ultimate decisions reduce the incentive for human effort.

A positive aspect of this effort pattern is that human operators can now focus more on troubleshooting and innovation (reducing type-2 errors) when technology

[7]For the official transcript released by the Federal Aviation Administration regarding the Northwest Airlines incident, see https://www.faa.gov/data_research/accident_incident/2009-10-23. For the preliminary investigation report on the Batik Air incident, refer to https://knkt.go.id/Repo/Files/Laporan/Penerbangan/2024/KNKT.24.01.02.04-Preliminary-Report.pdf.

[8]For news coverage on this issue, see https://www.businessinsider.com/tesla-driver-asleep-wheel-car-reached-70-mph-autopilot-times-2022-12 and https://www.nbcnews.com/news/us-news/tesla-driver-slept-car-was-going-over-80-mph-autopilot-n1267805, among others.

assumes final execution. A notable example from aviation is the US Airways Flight 1549 incident. After experiencing a bird strike shortly after takeoff from LaGuardia Airport in New York, the Airbus A320 lost both engines but successfully ditched in the Hudson River. The autopilot system (fly-by-wire) maintained the plane's basic stability (minimizing type-1 execution errors), allowing the pilots to concentrate on troubleshooting the problem and devising a solution within the extremely limited timeframe of three minutes.[9]

Anecdotally, recent developments in generative AIs (e.g., ChatGPT) enable software developers and artists to delegate more tedious final execution tasks (such as coding and video/image generation) to these tools. Consequently, humans can concentrate on the more innovative aspects of their work, such as algorithm design and concepts for original content.

4.2 Multiple Players and Effort Specialization

The insight from Proposition 5 extends to scenarios involving multiple layers of human decision-makers: early layers do not exert effort to reduce errors because early-stage errors are inconsequential; middle layers specialize in reducing the more relevant type-2 errors due to the relatively low-quality inputs; and final layers focus on minimizing type-1 errors as the input quality improves. To formalize these insights succinctly, I consider a simplified effort technology as follows.

Suppose there is a sequence of N ex-ante identical human players, each with a type-i error e_i, where $i = 1$ or 2. Each player can either use their status-quo technology, or, by incurring a cost c, reduce their type-1 error e_1 or type-2 error e_2 by Δ. I impose a natural parameter assumption that

$$e_2 - \Delta > e_1 > \Delta,$$

implying that even after effort is exerted to reduce type-2 error, it still dominates type-1 error, and the type-1 error remains positive after its reduction. Furthermore,

[9]In its submission to the National Transportation Safety Board regarding the accident investigation, Airbus explicitly stated that "[d]uring this time period the Aircraft was in the alpha protection mode which allowed the flight crew to remain focused on their priorities; conversely, if the Aircraft had been a non-fly-by-wire aircraft, the flight crew would have had to fly in and out of the stick shaker to maintain the desired descent profile." For full details, see https://data.ntsb.gov/Docket/Document/docBLOB?ID=40329236&FileExtension=.PDF&FileName=Airbus%20Submis
Master.PDF

to focus on the most generic case, I assume that the status-quo technology $\{e_1, e_2\}$ is useful in that it improves the probability of the correct outcome. Using Proposition 2 and (12), this assumption can be explicitly written as

$$p_0 < \frac{1-e_2}{1-e_2+e_1}. \tag{27}$$

As before, each player receives 1 unit of utility only if a correct action is ultimately adopted and optimally decides whether to exert effort and which type of error to reduce. I characterize the effort pattern in a Nash equilibrium, where no player has an incentive to deviate given the equilibrium effort profile of the other players.

Proposition 6 ***[Effort Specialization]*** *All Nash equilibria are characterized by two cutoffs $0 \leq N_1 \leq N_2 \leq N$, resulting in at most three effort regions. Players in the initial N_1 layers exert no effort; those between layers N_1+1 and N_2 specialize in reducing type-2 errors; and those in the final $N-N_2$ layers specialize in reducing type-1 errors.*

It is useful to be more explicit about how the equilibrium cutoffs N_1 and N_2 are determined. In the generic case when all three effort regions exist, the posterior probabilities can be explicitly calculated from equation (14) as follows:

$$p_j = \begin{cases} p^*_{(0)} - (e_2-e_1)^j \left(p^*_{(0)} - p_0\right) & j \leq N_1 \\ p^*_{(2)} - (e_2-e_1-\Delta)^{j-N_1} \left(p^*_{(2)} - p_{N_1}\right) & N_1 < j \leq N_2 \\ p^*_{(1)} - (e_2-e_1+\Delta)^{j-N_2} \left(p^*_{(1)} - p_{N_2}\right) & j > N_2 \end{cases}, \tag{28}$$

where $p^*_{(0)} \equiv \frac{1-e_2}{1-e_2+e_1}$, $p^*_{(1)} \equiv \frac{1-e_2}{1-e_2+e_1-\Delta}$, and $p^*_{(2)} \equiv \frac{1-e_2+\Delta}{1-e_2+\Delta+e_1}$ denote the invariant probabilities associated with the status-quo technology, and the ones after the reduction of type-1 or type-2 errors, respectively.[10]

In the early stages of the decision-making process, the incentive to exert effort is low because mistakes at these stages have limited consequences. As a result, initial

[10] All three regions exist when the initial probability p_0 is sufficiently small and $e_1 - \Delta < 1 - e_2$ holds. The second condition implies that reducing type-1 error is associated with higher invariant probabilities than reducing type-2 errors: $p^*_{(1)} > p^*_{(2)} > \max\{\frac{1}{2}, p^*_{(0)}\} > p_0$. Hence, the ordering of the matrices also satisfies the optimal integration rule specified in Proposition 2.

players do not exert effort until layer N_1, where

$$N_1 = \max\left\{j | c \leq (1 - p_{j-1})\,\Delta\,(e_2 - e_1 - \Delta)^{N_2 - j}\,(e_2 - e_1 + \Delta)^{N - N_2}\right\}. \tag{29}$$

The right-hand side of this inequality represents the difference in payoffs between not making effort and making effort to reduce type-2 errors.

The next group of players, after layer N_1, focus on reducing type-2 errors because the proposed action is less likely to be correct ($p_j \leq \frac{1}{2}$) making type-2 errors more relevant.

When the proposed action becomes more likely to be correct ($p_j > \frac{1}{2}$) in layer N_2, where

$$N_2 = \max\{j | p_j \leq \frac{1}{2}\}, \tag{30}$$

players shift their focus to the more relevant type-1 errors. The overall intuition follows and generalizes Proposition 5 to the case of multiple players.

It is worth noting that some of the three effort regions described in Proposition 6 can disappear. For example, if $N_2 = N$, then no players reduce type-1 errors. If $N_1 = 0$, then every player makes effort to reduce either type-1 or type-2 errors. Finally, if $N_2 \leq N_1$, then players in the initial N_1 layers do not make effort, those after layer N_1 specialize in reducing type-1 errors, and no players reduce type-2 errors.

Implications for Multi-layer Approval Process

Revisiting the application in Section 2.2.3 on multi-layer approval process, Proposition 6 provides insights into the division of labor among decision makers in different layers.

Initial layers, such as departmental or school committees, often dedicate substantial effort to screening promotion cases for merit, addressing errors and rectifying misjudgments from earlier rounds (reducing type-2 errors). At this stage, inputs are relatively noisy—a candidate may be misevaluated by letter writers or other initial reviewers. Consequently, the relevance of addressing type-2 errors is high.

In contrast, higher-level decision-makers, such as provosts and presidents, rarely overturn school recommendations. Their focus is instead on the efficient execution of decisions: ensuring procedural compliance, minimizing administrative burdens, and facilitating smooth outcomes for both successful and unsuccessful candidates. This includes promoting successful candidates effectively and dismissing unsuccessful ones in a way that avoids adverse consequences for the university and the individuals

involved. At this stage, recommendations from earlier layers are more likely to be accurate, shifting the emphasis to execution and reducing type-1 errors.

A similar division of focus is observed in many other multi-layer approval processes, such as budget reviews and student admissions. Lower-tier agents prioritize error reduction, while higher-tier decision-makers, who hold ultimate authority, emphasize procedural fidelity and executional accuracy.

5 Two Applications

5.1 Integration of Human and Generative AI

This section applies the framework to analyze the integration of humans and generative AI in decision-making processes. By adding structure to the decision matrix, the model accommodates several key features of generative AI technologies.

First, generative AI can produce novel outcomes without requiring extensive human input. These outcomes may occasionally surpass human capabilities (e.g. AlphaGo's triumph over human players) or result in mistakes, such as AI hallucinations (e.g., producing erroneous or misleading content). Second, AI can handle a large volume of tasks, significantly reducing the effort required from humans to make active decisions.[11]

To capture these features, suppose there are three outcomes with associated payoffs, $\mathbf{U} = \{S, 1, 0\}$:

1. The outcome of 0 represents a mistake.

2. The payoff of 1 corresponds to a routine correct outcome, which can be delivered by either humans or AI.

3. The special outcome with a payoff of S represents a unique result achievable only by AI. The value S can be greater or smaller than 1, reflecting the creative potential of AI, which may exceed or fall short of human capabilities.

The prior distribution of outcomes consists of either correct results or mistakes (i.e., no special outcomes) and is given by $\mathbf{P}_0 = (0, p_0, 1 - p_0)$.

[11] See Banh and Strobel (2023) for a discussion on the key features and challenges associated with generative AI.

If humans actively make decisions, their effect is described by the decision matrix:

$$\mathcal{M}_H = \begin{pmatrix} 1 - e_{1,H} & 0 & e_{1,H} \\ 0 & 1 - e_{1,H} & e_{1,H} \\ 0 & 1 - e_{2,H} & e_{2,H} \end{pmatrix}.$$

where $e_{1,H}$ represents the human type-1 error—the possibility of failing to maintain a good status-quo outcome (either a prior or AI-generated result) and $e_{2,H}$ represents the human type-2 error—the possibility of failing to correct a mistake.

To model the cost of human decisions, I assume that "no change" is always an option at no cost. This scenario is mathematically represented by the identity matrix $\mathcal{I}$. Active decision making through $\mathcal{M}_H$, however, is associated with a cost $c(\pi_H)$ where π_H indicates the fraction of tasks (or equivalently, probability) for which human actively make decisions. No change is made to the remaining $1-\pi_H$ fraction of tasks.

Combining these two aspects, the human decision matrix is expressed as:

$$(1 - \pi_H)\,\mathcal{I} + \pi_H \mathcal{M}_H.$$

Unlike active human decisions, AI's decisions do not incur additional costs.[12] In addition, AI has the unique ability to achieve special outcomes with a probability s_{AI}. However, similar to human, AI is also prone to type-1 ($e_{1,AI}$) and type-2 ($e_{2,AI}$) errors. This is captured by the AI decision matrix:

$$\mathcal{M}_{AI} = \begin{pmatrix} a & b & c \\ s_{AI} & 1 - s_{AI} - e_{1,AI} & e_{1,AI} \\ s_{AI} & 1 - s_{AI} - e_{2,AI} & e_{2,AI} \end{pmatrix},$$

where the probabilities $\{a, b, c\}$, conditional on the special outcome, are arbitrary and irrelevant because AI does not encounter special outcomes as inputs in the model.[13]

Determining whether humans or AI should make the final decision depends on the

[12]The large fixed cost of adopting AI is sunk and therefore does not appear in the optimal integration problem.

[13]Recall that special outcomes are defined as outcomes only achievable by AI.

following comparison:

$$\begin{cases} \max_{\pi_H} \mathbf{P}_0 \left((1-\pi_H)\mathcal{I} + \pi_H \mathcal{M}_H \right) \mathcal{M}_{AI} \mathbf{U}' - c(\pi_H), & \text{if AI makes the final decision,} \\ \max_{\pi_H} \mathbf{P}_0 \mathcal{M}_{AI} \left((1-\pi_H)\mathcal{I} + \pi_H \mathcal{M}_H \right) \mathbf{U}' - c(\pi_H), & \text{if human makes the final decision.} \end{cases} \tag{31}$$

It is useful to note that the optimal effort level, π_H, generally depends on the sequence of integration (i.e., whether humans act before or after AI). After simple manipulations, the payoff functions can be rewritten as:

$$\begin{cases} \max_{\pi_H} \pi_H \mathbf{P}_0 \left(\mathcal{M}_H - \mathcal{I} \right) \mathcal{M}_{AI} \mathbf{U}' + \mathbf{P}_0 \mathcal{M}_{AI} \mathbf{U}' - c(\pi_H), & \text{if AI makes the final decision,} \\ \max_{\pi_H} \pi_H \mathbf{P}_0 \mathcal{M}_{AI} \left(\mathcal{M}_H - \mathcal{I} \right) \mathbf{U}' + \mathbf{P}_0 \mathcal{M}_{AI} \mathbf{U}' - c(\pi_H), & \text{if human makes the final decision.} \end{cases} \tag{32}$$

Define three effort levels:

1. $\pi^*_{H,1}$: the optimal effort level when AI makes the final decision (the first row in (32));
2. $\pi^*_{H,2}$: the optimal effort level when human makes the final decision (the second row in (32));
3. $\pi^*_{H,0}$: the optimal effort level without AI:

$$\pi^*_{H,0} \equiv \arg\max_{\pi_H} \pi_H \mathbf{P}_0 \left(\mathcal{M}_H - \mathcal{I} \right) \mathbf{U}' + \mathbf{P}_0 \mathbf{U}' - c(\pi_H). \tag{33}$$

I can now characterize the optimal integration of human and AI, and compare human effort level in different scenarios.

Proposition 7 *It is optimal for the AI to make the final decision if and only if*

$$\frac{1-e_{2,AI}}{e_{1,AI}} \left[\frac{s_{AI}}{1-e_{2,AI}} (S-1) + 1 \right] \geq \frac{1-e_{2,H}}{e_{1,H}}, \tag{34}$$

Under this optimal integration, human effort is higher compared to the suboptimal integration:

$$\pi^*_{H,1} \geq \pi^*_{H,2}.$$

Furthermore, the presence of AI reduces human effort level compared to the scenario

without AI:

$$\pi^*_{H,1} \leq \pi^*_{H,0}.$$

To understand this result, first note that the integration rule (34) closely relates to those characterized by Propositions 1 and 2. In fact, when AI does not generate any special outcome $s_{AI} = 0$ or if the special outcome is not distinctive ($S = 1$), condition (34) degenerates to the case of binary outcomes: comparing invariant probabilities ($\frac{1-e_2}{1-e_2+e_1}$ which is equivalent to $\frac{1-e_2}{e_1}$) for both human and AI.

AI's ability to generate outcomes superior to the correct action ($S > 1$) enhances its effective "quality," making it more likely to serve as the final decision maker. Conversely, when $S < 1$, such AI is more likely to be integrated before human, acting as an assistive tool.

For AI to take final control, it is not necessary for its error detection capability to surpass that of humans (i.e., $e_{2,AI} < e_{2,H}$). The critical factor is that the AI must not generate a significant amount of new mistakes (i.e., its type-1 error $e_{1,AI}$ must not be too large). Mathematically, in condition (34), type-1 error appears in the denominator, amplifying its importance. Put differently, a high probability of AI hallucination (high type-1 errors) significantly constrains its feasibility as the final decision maker.

Next, I turn to the effort comparison in Proposition 7. Revisiting (32), when humans "do nothing," the AI-only decision outcome is represented by $\mathbf{P}_0\mathcal{M}_{AI}\mathbf{U}'$, regardless of the integration sequence. The marginal benefit of involving humans lies in their active decision-making, which, with probability π_H, yields $(\mathcal{M}_H - \mathcal{I})$ in the corresponding layer. This marginal benefit is captured by the coefficient of π_H in both cases of (32). The significance of this marginal benefit determines both the level of human effort and the optimal integration sequence. Consequently, the optimal integration sequence also motivates higher human effort compared with the suboptimal integration sequence.

It is worth noting that this insight is independent of the specific forms of $\mathcal{M}_H$ and $\mathcal{M}_{AI}$. However, it does rely on the assumption that the benchmark action is "doing nothing," represented by the identity matrix $\mathcal{I}$. Suppose human effort modifies the

decision matrix from a generic $\mathcal{M}_H^{ne}$ to $\mathcal{M}_H^e$, then (32) becomes

$$\begin{cases} \max_{\pi_H} \pi_H \mathbf{P}_0 \left(\mathcal{M}_H^e - \mathcal{M}_H^{ne}\right) \mathcal{M}_{AI} \mathbf{U}' + \mathbf{P}_0 \mathcal{M}_H^{ne} \mathcal{M}_{AI} \mathbf{U}' - c(\pi_H), & \text{if AI makes the final decision,} \\ \max_{\pi_H} \pi_H \mathbf{P}_0 \mathcal{M}_{AI} \left(\mathcal{M}_H^e - \mathcal{M}_H^{ne}\right) \mathbf{U}' + \mathbf{P}_0 \mathcal{M}_{AI} \mathcal{M}_H^{ne} \mathbf{U}' - c(\pi_H), & \text{if human makes the final decision.} \end{cases} \tag{35}$$

Clearly, the optimal integration depends not only on the marginal effect effect $(\mathcal{M}_H^e - \mathcal{M}_H^{ne})$ but also how the benchmark no-effort matrix $\mathcal{M}_H^{ne}$ integrates with AI's decision matrix $\mathcal{M}_{AI}$.

The presence of AI reduces the necessity of human effort, aligning broadly with the general intuition in Proposition 5. The underlying mechanism is that the marginal benefit of active decision-making decreases due to the presence of another subsequent layer.

5.2 Multi-layer Delegation

The framework developed in this paper extends the classic delegation problem by Aghion and Tirole (1997) to a multi-layer setting.

Consider a principal and $N \geq 2$ agents, indexed by n. Following Aghion and Tirole (1997), the principal needs to adopt one project out of many ex-ante identical projects. The n-th agent's preferred project is labeled as the n-th project. Additionally, there is an $N+1$-th project, yielding a payoff of 0 to both the principal and the agents, which can be interpreted as doing nothing. Apart from these $N+1$ "relevant" projects, there are also disastrous projects that generate sufficiently negative payoffs. The identities of the projects are unknown to all players ex-ante, except for the $N+1$-th project (doing nothing). Thus, in the absence of additional information, the $N+1$-th project should be adopted to avoid the risk of disastrous projects. Consequently, only the first $N+1$ projects can be chosen in equilibrium.

The n-th agent learns the identity of their preferred project with probability q_n. To account for the possibility of incorrect information and mistakes, I assume that with probability $q_n' \leq 1 - q_n$, the agent erroneously overrules the previously proposed project and instead proposes doing nothing (i.e., the $N+1$-th project). With the residual probability $1 - q_n - q_n'$, the agent has no information and retains the proposed project choice. The agent's preferred choice generates a payoff of U_n to the principal. Each agent is therefore represented by a triplet (q_n, q_n', U_n), which measures the agent's

skill level (q_n), propensity for mistakes (q'_n), and preference alignment or loyalty (U_n).

The multi-layer delegation game unfolds as follows. The principal selects a delegation sequence $\sigma(n)$, assigning the n-th agent in layer $i = \sigma(n)$. The initial proposal is doing-nothing (project $N+1$), i.e., the initial probability distribution $\mathbf{P}_0 = (0, ..., 0, 1) \in \mathbb{R}^{N+1}$ over the $N+1$ relevant projects. The agent in layer i maintains the proposed project from the previous layer $i-1$ if he has no information (with probability $1 - q_n - q'_n$). In addition, with probability q_n, the agent correctly identifies and proposes his preferred project n. Finally, with probability q'_n the agent mistakenly overrules prior recommendations and proposes project $N+1$. Hence, the agent's decision is summarized by the following $(N+1) \times (N+1)$ matrix

$$\mathcal{M}_n = \begin{pmatrix} 1-q_n-q'_n & 0 & 0 & q_n & 0 & 0 & 0 & q'_n \\ 0 & 1-q_n-q'_n & 0 & q_n & 0 & 0 & 0 & q'_n \\ 0 & 0 & ... & q_n & ... & ... & ... & ... \\ 0 & 0 & 0 & 1-q'_n & 0 & 0 & 0 & q'_n \\ 0 & 0 & 0 & q_n & 1-q_n-q'_n & 0 & 0 & q'_n \\ & & & & & ... & & \\ 0 & 0 & 0 & q_n & 0 & 0 & 1-q_n-q'_n & q'_n \\ 0 & 0 & 0 & q_n & 0 & 0 & 0 & 1-q_n \end{pmatrix},$$

where the n-th and the $N+1$-th columns contain the (off-diagonal) entries q_n and q'_n, respectively.

The principal optimally chooses the delegation sequence $\sigma(n)$ to maximize the expected payoff as described in (5), where $\mathbf{U} \equiv (U_1, U_2, ..., U_N, 0)$.

Proposition 8 *Suppose there are N agents, each associated with a triplet (q_n, q'_n, U_n). The optimal delegation order σ^* is given by*

$$\sigma^*(n) = 0 \text{ if and only if } U_n < 0,$$

and

$$\sigma^*(n_1) > \sigma^*(n_2) > 0 \text{ if and only if } \frac{U_{n_1}}{1 + \frac{q'_{n_1}}{q_{n_1}}} > \frac{U_{n_2}}{1 + \frac{q'_{n_2}}{q_{n_2}}}. \tag{36}$$

In words, only agents with $U_n \geq 0$ are included in the optimal delegation sequence, and they are arranged in ascending order of $\frac{U_n}{1+\frac{q'_n}{q_n}}$. Similar to the role of invariant

probabilities p_n^* in Theorem 2, the "quality" of agents in this delegation application is measured by $\frac{U_n}{1+\frac{q_n'}{q_n}}$. Intuitively, more loyal agents (higher U_n) and more skilled agents (higher q_n) and agents who make fewer mistakes (lower q_n') are more likely to be delegated the responsibility of deciding on the project.

This application nests the two-party principal-agent delegation problem in the classic Aghion-Tirole (1997) framework as a special case: $N = 2$, $U_1 = 1$, and $U_2 = \alpha < 1$. Essentially, agent 1 is the principal, sharing the same preferred action whose payoff is normalized to $U_1 = 1$, and agent 2's preferred action generates a lower payoff. Having agent 1 act first ($\mathcal{M}_1\mathcal{M}_2$) reflects effective delegation to the agent (or "A-formal authority" in Aghion-Tirole's terminology), while having agent 2 act first ($\mathcal{M}_2\mathcal{M}_1$) reflects the principal retaining ultimate authority (or "P-formal authority"). Consistent with their findings, the principal is more likely to delegate when the agent is more loyal (higher U_2), more skilled (higher q_2), and less prone to mistakes (lower q_1). Proposition 8 extends the basic Aghion-Tirole result to a multi-layer setting, offering a clear order for arranging heterogeneous agents.

An intriguing observation is that when agents do not make mistakes ($q_n' = 0$), the optimal delegation order depends only on loyalty (U_n) and not on skill (q_n). Put differently, more loyal agents are more likely to become ultimate decision-makers, regardless of their skill. Intuitively, this arrangement allows more loyal agents to select projects closer to the principal's preferences. Furthermore, it does not compromise the contributions of skilled but less loyal agents, whose preferred projects are still adopted if the more loyal agents remain uninformed about their own preferences.

6 Technical Extension: Higher type-1 error ($e_1 > e_2$)

Throughout this work, in the case of binary actions ($A = 2$), I have assumed that correcting a mistake is more difficult than maintaining a correct action, i.e., $e_2 > e_1$, which is arguably the more empirically relevant case. In this section, I consider the opposite scenario where maintaining the correct action is more difficult, i.e., $e_1 > e_2$. Such a technology delivers better outcomes when the prior probability is worse. Mathematically, this is evident from (15), where p_i is negatively related to p_{i-1}. Furthermore, (14) shows that the prior and posterior probabilities p_{i-1} and p_i oscillate around the invariant probability p_i^*. Intuitively, when a technology is better at correcting errors than maintaining correctness, feeding the technology a mistake

results in a superior outcome.

Consequently, in contrast to Proposition 2 and Corollary 1, the optimal integration in this case may no longer be "monotonic." It might be optimal to initially degrade the quality of the proposed action and then apply a technology with a lower type-2 error to correct these mistakes, thereby achieving a better outcome. Moreover, it is possible for "bad" technologies with invariant probabilities lower than the initial p_0 to be utilized in the optimal integration.

Consider the following simple and extreme example with three technologies:

$$\mathcal{M}_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 1 \end{pmatrix}, \quad \mathcal{M}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \text{and } \mathcal{M}_3 = \begin{pmatrix} 0.1 & 0.9 \\ 1 & 0 \end{pmatrix},$$

with the initial probability $p_0 = 0.7$.

- The first technology ($\mathcal{M}_1$) never corrects existing errors, introduces 50% new errors, and therefore has an invariant probability $p_1^* = \frac{1-1}{1-1+0.5} = 0$.
- The second technology ($\mathcal{M}_2$) exchanges the correct and wrong actions, yielding an invariant probability $p_2^* = \frac{1-0}{1-0+1} = 0.5$.
- The third technology ($\mathcal{M}_3$) has a slightly lower type-1 error of 90%, resulting in an invariant probability $p_3^* = \frac{1-0}{1-0+0.9} = 0.53$.

Notably, both $\mathcal{M}_2$ and $\mathcal{M}_3$ have higher type-1 errors ($e_1 > e_2$). Despite all three technologies having invariant probabilities lower than p_0, the optimal integration sequence is $\mathcal{M}_2\mathcal{M}_1\mathcal{M}_3$, delivering a final correct probability of 86.5%.

Examining the effect after each layer, following $\mathcal{M}_2$, the posterior probability becomes $1 - p_0 = 30\%$. After the second layer ($\mathcal{M}_2\mathcal{M}_1$), the posterior further deteriorates to 15% as $\mathcal{M}_1$ introduces another 50% new mistakes. Finally, $\mathcal{M}_3$ corrects all existing mistakes while maintaining 10% of the correct outcomes, resulting in a final probability of $1 - 15\% + 0.1 * 15\% = 86.5\%$.

As discussed, when $e_1 > e_2$, the prior and posterior probabilities oscillate around the invariant probability. This oscillation is evident in $\mathcal{M}_2$ ($70\% \to 30\%$ after layer 1) and $\mathcal{M}_3$ ($15\% \to 86.5\%$ after layer 3). Due to this oscillation, even a technology that solely generates errors ($\mathcal{M}_1$) may prove useful in the optimal integration.

These complexities make it challenging to characterize a general rule for optimal integration, leaving room for future research to advance our understanding in this

direction.

7 Conclusion

This paper constructs a framework based on Markov matrices to analyze the optimal sequential integration of different decision-making technologies. In the binary-action case, technologies differ in their type-1 errors (rejecting a correct action) and type-2 errors (accepting a wrong action). The analysis demonstrates that applying technologies in ascending order of their associated invariant probabilities yields the best ultimate decision. However, when there are multiple payoff-relevant states or more than two actions, such a one dimensional integration rule does not generally exist.

The study also reveals that human effort to reduce errors increases throughout the decision-making process, with efforts to minimize type-1 errors accelerating more rapidly than efforts to minimize type-2 errors. This finding suggests that early decision-makers should prioritize reducing type-2 errors, while final decision-makers should focus on reducing type-1 errors.

I provide two applications of the framework. In the context of human-AI integration, the model provides a decision rule for assigning final decision authority to a generative AI. Specifically, the AI is more likely to be entrusted with ultimate decision-making when its error rates are low, and its creative abilities are high. Moreover, the presence of AI reduces the necessity for human decision-making effort. In the context of multi-layer delegation, the model extends the classic Aghion and Tirole (1997) result, introducing a metric to rank agents based on their skill, loyalty, and propensity for errors. Additional applications include automation design, multi-layer approval processes, and loan screening.

Finally, the framework developed here can be further applied to study other layered structures. For example, Zhong (2023) employs a similar framework to examine the market structure of intermediation chains. Another promising area of application lies in labor hierarchies and job design. In the specific case of human-AI integration, numerous questions remain unanswered. For instance, in autonomous driving, how should rewards or liabilities be allocated between the driver and the car manufacturer? The allocation could depend on the interim outcomes of decisions—for example, whether the human corrects the AI's mistakes or vice versa. Additionally, there may be multiple drivers sharing the same autonomous driving system. Ad-

dressing these questions could provide valuable insights for regulation and insurance design in this market.

References

[1] Aghion, Philippe, and Jean Tirole. "Formal and real authority in organizations." Journal of political economy 105.1 (1997): 1-29.

[2] Banerjee, Abhijit V. "A simple model of herd behavior." The quarterly journal of economics 107.3 (1992): 797-817.

[3] Banh, Leonardo, and Gero Strobel. "Generative artificial intelligence." Electronic Markets 33.1 (2023): 63.

[4] Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. "A theory of fads, fashion, custom, and cultural change as informational cascades." Journal of political Economy 100.5 (1992): 992-1026.

[5] Calvo, Guillermo A., and Stanislaw Wellisz. "Hierarchy, ability, and income distribution." Journal of political Economy 87.5, Part 1 (1979): 991-1010.

[6] Chen, Cheng. "Management quality and firm hierarchy in industry equilibrium." American Economic Journal: Microeconomics 9.4 (2017): 203-244.

[7] Cong, Lin William, and Yizhou Xiao. "Information cascades and threshold implementation: Theory and an application to crowdfunding." The Journal of Finance 79.1 (2024): 579-629.

[8] Dasgupta, Amil, and Ernst G. Maug. "Delegation Chains." Available at SSRN 3952744 (2021).

[9] Garicano, Luis. "Hierarchies and the Organization of Knowledge in Production." Journal of political economy 108.5 (2000): 874-904.

[10] Glode, Vincent, and Christian Opp. "Asymmetric information and intermediation chains." American Economic Review 106.9 (2016): 2699-2721.

[11] Glode, Vincent, Christian C. Opp, and Xingtan Zhang. "On the efficiency of long intermediation chains." Journal of Financial Intermediation 38 (2019): 11-18.

[12] He, Zhiguo, and Jian Li. Intermediation via credit chains. No. w29632. National Bureau of Economic Research, 2022.

[13] Huang, Xin, Gengsheng Qin, and Yixin Fang. "Optimal combinations of diagnostic tests based on AUC." Biometrics 67.2 (2011): 568-576.

[14] Johnson, Norman L. "Sequential analysis: A survey." Journal of the Royal Statistical Society: Series A (General) 124.3 (1961): 372-411.

[15] Kornecki, Andrew J., and Kimberley Hall. "Approaches to assure safety in fly-by-wire systems: Airbus vs. boeing." IASTED Conf. on Software Engineering and Applications. 2004.

[16] Qian, Yingyi. "Incentives and loss of control in an optimal hierarchy." The review of economic studies 61.3 (1994): 527-544.

[17] Saaty, Thomas L. "A scaling method for priorities in hierarchical structures." Journal of mathematical psychology 15.3 (1977): 234-281.

[18] Stroock, Daniel W. An introduction to Markov processes. Vol. 230. Springer Science & Business Media, 2013.

[19] Su, John Q., and Jun S. Liu. "Linear combinations of multiple diagnostic markers." Journal of the American Statistical Association 88.424 (1993): 1350-1355.

[20] Wald, Abraham. "Sequential Tests of Statistical Hypotheses". Annals of Mathematical Statistics. 16.2 (1945): 117–186.

[21] Zhong, Hongda. "Market structure of intermediation." Available at SSRN 4185231 (2023).

Appendix

Proof of Proposition 1: Note that both sides of (16) are linear functions in p_0. Hence, I only need to establish the inequality at the two boundaries $p_0 = 0$ and $p_0 = p_1^*$, and then the result holds for all $p_0 \leq p_1^*$.

First, consider $p_0 = p_1^*$. Since $p_1^* < p_2^*$, it follows from (14) that

$$\left(\begin{array}{cc} p_1^* & 1-p_1^* \end{array} \right) \mathcal{M}_2 \left(\begin{array}{c} 1 \\ 0 \end{array} \right) \in (p_1^*, p_2^*) . \tag{37}$$

Apply (14) again, the lower bound of p_1^* in (37) implies that

$$\left(\begin{array}{cc} p_1^* & 1-p_1^* \end{array} \right) \mathcal{M}_2 \mathcal{M}_1 \left(\begin{array}{c} 1 \\ 0 \end{array} \right) < \left(\begin{array}{cc} p_1^* & 1-p_1^* \end{array} \right) \mathcal{M}_2 \left(\begin{array}{c} 1 \\ 0 \end{array} \right) = \left(\begin{array}{cc} p_1^* & 1-p_1^* \end{array} \right) \mathcal{M}_1 \mathcal{M}_2 \left(\begin{array}{c} 1 \\ 0 \end{array} \right) .$$

Next, consider $p_0 = 0$. Our objective (16) becomes

$$(1-e_{1,2})(1-e_{2,1}) + (1-e_{2,2})e_{2,1} > (1-e_{1,1})(1-e_{2,2}) + (1-e_{2,1})e_{2,2}$$

which is equivalent to

$$(1-e_{1,2}-e_{2,2})(1-e_{2,1}) > (1-e_{1,1}-e_{2,1})(1-e_{2,2})$$

which is in turn equivalent to

$$e_{1,2}(1-e_{2,1}) < e_{1,1}(1-e_{2,2}) \Leftrightarrow \frac{e_{1,1}}{1-e_{2,1}} > \frac{e_{1,2}}{1-e_{2,2}} \Leftrightarrow p_2^* > p_1^* .$$

Finally, suppose $p_1^* = p_2^*$. Then matrices $\mathcal{M}_i$ can be simultaneously diagonalized as follows

$$\left(\begin{array}{cc} 1 & 1-p_i^* \\ 1 & -p_i^* \end{array} \right) = \left(\begin{array}{cc} p_i^* & 1-p_i^* \\ 1 & -1 \end{array} \right)^{-1}$$

$$\mathcal{M}_i = \left(\begin{array}{cc} 1 & 1-p_i^* \\ 1 & -p_i^* \end{array} \right) \left(\begin{array}{cc} 1 & 0 \\ 0 & e_{2,i}-e_{1,i} \end{array} \right) \left(\begin{array}{cc} p_i^* & 1-p_i^* \\ 1 & -1 \end{array} \right) .$$

Hence, the two matrices commute, and the ordering is irrelevant. ■

Proof of Proposition 2: I first establish the following lemma.

Lemma 1 *For any integration σ of N technologies, the final probability p_N is strictly increasing in p_0.*

Proof of Lemma 1: It is clear from (14) that since $e_{2,n} > e_{1,n}$, the posterior probability p_i is strictly increasing in the prior probability p_{i-1}. Mathematical induction immediately implies that it also strictly increases with the prior probability p_0. ■

Now I prove the proposition. First, I show all technologies with $p_n^* < p_0$ should be excluded from the optimal integration. Suppose otherwise, the nth technology $\mathcal{M}_n$ and $\sigma^*(n) = N_0$ is the first technology in the integration with $p_n^* < p_0$. Consequently, all technologies in the integration before N_0 ($i < N_0$) have $p^*_{\sigma^{*-1}(i)} > p_0$. From (14), we know that the posterior probability in the integration after applying $N_0 - 1$ technologies is higher than p_0. Hence, removing technology $\mathcal{M}_n$ strictly increases the posterior probability after N_0 technologies. Lemma 1 then implies that the final probability is also strictly higher, contradicting with the efficiency of σ^*. Hence, all technologies in the optimal integration must feature invariant probabilities higher than p_0.

In addition, the optimal integration must include all technologies with invariant probability greater than p_0. Otherwise, suppose $\mathcal{M}_n$ is an excluded technology with $p_n^* > p_0$, i.e., $\sigma^*(n) = 0$. Then adding it as the first technology in the integration ($\hat{\sigma}(n) = 1$ and $\hat{\sigma}(i) = \sigma^*(i) + 1$ for all $\sigma^*(i) > 0$) strictly improves the posterior probability p_1 as well as the final probability per Lemma 1.

Next, I show that the optimal integration must feature a sequence of technologies with weakly increasing invariant probabilities. Suppose otherwise that N_0 and $N_0 + 1$ are two adjacent technologies in the optimal integration σ^* such that $p^*_{\sigma^{*-1}(N_0)} > p^*_{\sigma^{*-1}(N_0+1)}$. Consider the posterior probability p_{N_0-1} after applying the first $N_0 - 1$ technologies. There are two possibilities.

If $p_{N_0-1} \geq p^*_{\sigma^{*-1}(N_0+1)}$, then the posterior probability p_{N_0} after N_0 must lie between p_{N_0-1} and $p^*_{\sigma^{*-1}(N_0)}$. Therefore, the posterior $p_{N_0} > p^*_{\sigma^{*-1}(N_0+1)}$ and it is optimal to remove the $N_0 + 1$th technology.

If $p_{N_0-1} < p^*_{\sigma^{*-1}(N_0+1)}$, then Proposition 1 implies that switching the order between the N_0th and $N_0 + 1$th technology strictly improves the posterior probability.

Both cases contradict with σ^* being the optimal integration, and hence the proposition. ■

Proof of Proposition 3: For any $\mathbf{p}_0^{(t)} \equiv (p_0, 1-p_0)$, calculate the difference in payoff between the two integrations $p_2 \equiv \mathbf{p}_0^{(t)} \mathcal{M}_1^{(t)} \mathcal{M}_2^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\hat{p}_2 = \mathbf{p}_0^{(t)} \mathcal{M}_2^{(t)} \mathcal{M}_1^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. First, consider the former integration p_2. Iterate (14),

$$\begin{aligned} p_2^* - p_2 &= (e_{2,2} - e_{1,2})(p_2^* - p_1^* + p_1^* - p_1) \\ &= (e_{2,2} - e_{1,2})(p_2^* - p_1^*) + (e_{2,2} - e_{1,2})(e_{2,1} - e_{1,1})(p_1^* - p_0). \end{aligned}$$

Symmetry implies that, for the alternative integration $\hat{p}_2$, one has

$$p_1^* - \hat{p}_2 = (e_{2,1} - e_{1,1})(p_1^* - p_2^*) + (e_{2,2} - e_{1,2})(e_{2,1} - e_{1,1})(p_2^* - p_0).$$

Taking the difference

$$\begin{aligned} p_2 - \hat{p}_2 &= (p_2^* - p_1^*)\left[1 - (e_{2,2} - e_{1,2}) - (e_{2,1} - e_{1,1}) + (e_{2,2} - e_{1,2})(e_{2,1} - e_{1,1})\right] \\ &= (p_2^* - p_1^*)\left[1 - (e_{2,2} - e_{1,2})\right]\left[1 - (e_{2,1} - e_{1,1})\right]. \end{aligned}$$

Taking all types together, the payoff difference (18) becomes

$$\mathbf{P}_0 \mathcal{M}_1 \mathcal{M}_2 \mathbf{U}' - \mathbf{P}_0 \mathcal{M}_2 \mathcal{M}_1 \mathbf{U}' = \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) q_t (p_2^{(t)*} - p_1^{(t)*}) \left[1 - (e_{2,2}^{(t)} - e_{1,2}^{(t)})\right] \left[1 - (e_{2,1}^{(t)} - e_{1,1}^{(t)})\right]$$

Using the definition of $p_i^{(t)*}$, one can easily verify that $\mathbf{P}_0 \mathcal{M}_1 \mathcal{M}_2 \mathbf{U}' \geq \mathbf{P}_0 \mathcal{M}_2 \mathcal{M}_1 \mathbf{U}'$ is equivalent to

$$\sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) q_t \left[\frac{1}{p_1^{(t)*}} - \frac{1}{p_2^{(t)*}}\right] \left(1 - e_{2,2}^{(t)}\right) \left(1 - e_{2,1}^{(t)}\right) \geq 0.$$

Together with

$$\frac{1}{p_i^{(t)*}} = \frac{e_{1,i}^{(t)}}{1 - e_{2,i}^{(t)}} + 1,$$

The above condition becomes

$$\sum_{t=1}^{T}\left(\overline{u}^{(t)}-\underline{u}^{(t)}\right)q_t\left[\frac{e_{1,1}^{(t)}}{1-e_{2,1}^{(t)}}-\frac{e_{1,2}^{(t)}}{1-e_{2,2}^{(t)}}\right]\left(1-e_{2,2}^{(t)}\right)\left(1-e_{2,1}^{(t)}\right)\geq 0.$$

After manipulation, it becomes

$$\sum_{t=1}^{T}\left(\overline{u}^{(t)}-\underline{u}^{(t)}\right)q_t\left[e_{1,1}^{(t)}\left(1-e_{2,2}^{(t)}\right)-e_{1,2}^{(t)}\left(1-e_{2,1}^{(t)}\right)\right]\geq 0,$$

establishing conditions (19) and (20). ■

Proof of Proposition 5: Compare the two cases when the player operates in layer j versus layer $j+1$. The posterior probability after layer $j-1$ is p_{j-1}. For notational convenience, denote by $\{e_{i,k}|i=1,2 \text{ and } k=1,2,...,N\}$ the error probabilities associated with the remaining $N-1$ technologies, and $e_{i,j}=e_{i,j+1}$ is same technology in either layer j or $j+1$ depending where the player operates. If the player is in layer j, the first order conditions (23) and (24) can be rewritten as

$$-p_{j-1}h_1'(f_{1,j}^*)\Pi_{i=j+1}^{N}\left(e_{2,i}-e_{1,i}\right)=c'(f_{1,j}^*), \tag{38}$$

and

$$-\left(1-p_{j-1}\right)h_2'(f_{2,j}^*)\Pi_{i=j+1}^{N}\left(e_{2,i}-e_{1,i}\right)=c'(f_{2,j}^*). \tag{39}$$

If the machine is the first layer, then the player's effort levels $f_{1,j+1}^*$ and $f_{2,j+1}^*$ are given by

$$-\left[p_{j-1}\left(1-e_{1,j}\right)+\left(1-p_{j-1}\right)\left(1-e_{2,j}\right)\right]h_1'(f_{1,j+1}^*)\Pi_{i=j+2}^{N}\left(e_{2,i}-e_{1,i}\right)=c'(f_{1,j+1}^*), \tag{40}$$

and

$$-\left[p_{j-1}e_{1,j}+\left(1-p_{j-1}\right)e_{2,j}\right]h_2'(f_{2,j+1}^*)\Pi_{i=j+2}^{N}\left(e_{2,i}-e_{1,i}\right)=c'(f_{2,j+1}^*). \tag{41}$$

Corollary 1 implies that the posterior probabilities p_i are increasing in i. Comparing (38) and (40), I have

$$\frac{c_1'(f_{1,j}^*)}{-h_1'(f_{1,j}^*)}=p_{j-1}\Pi_{i=j+1}^{N}\left(e_{2,i}-e_{1,i}\right)>p_j\Pi_{i=j+2}^{N}\left(e_{2,i}-e_{1,i}\right)=\frac{c_1'(f_{1,j+1}^*)}{-h_1'(f_{1,j+1}^*)}.$$

Since $\frac{c_1'(f)}{-h_1'(f)}$ is an increasing function in f, it follows that $f_{1,j+1}^* > f_{1,j}^*$.

Next, I show that $f_{2,j+1}^* > f_{2,j}^*$ also holds. Note that since $e_{i,j}$ and $e_{i,j+1}$ feature the same technology (the one being swapped with human), we have

$$p_{j-1}e_{1,j} + (1-p_{j-1})\,e_{2,j} > (1-p_{j-1})\,(e_{2,j} - e_{1,j}) = (1-p_{j-1})\,(e_{2,j+1} - e_{1,j+1})\,,$$

Together with (39) and (41), I have

$$\frac{-h_1'(f_{1,j}^*)}{-h_2'(f_{2,j+1}^*)} > \frac{-h_1'(f_{1,j}^*)}{-h_2'(f_{2,j+1}^*)} \frac{(1-p_{j-1})\,(e_{2,j+1} - e_{1,j+1})}{[p_{j-1}e_{1,j} + (1-p_{j-1})\,e_{2,j}]} = \frac{c_2'(f_{2,j}^*)}{c_2'(f_{2,j+1}^*)}.$$

Hence, the increasing monotonicity of $\frac{c_2'(f)}{-h_2'(f)}$ again implies that $f_{2,j+1}^* > f_{2,j}^*$.

Finally, to establish (26), observe that

$$\frac{c_1'(f_{1,j+1}^*)}{-h_1'(f_{1,j+1}^*)} \Big/ \frac{c_1'(f_{1,j}^*)}{-h_1'(f_{1,j}^*)} = \frac{p_j}{p_{j-1}\,(e_{2,j+1} - e_{1,j+1})}$$

and

$$\frac{c_2'(f_{2,j+1}^*)}{-h_2'(f_{2,j+1}^*)} \Big/ \frac{c_2'(f_{2,j}^*)}{-h_2'(f_{2,j}^*)} = \frac{1-p_j}{(1-p_{j-1})\,(e_{2,j+1} - e_{1,j+1})}.$$

The fact that the technologies are optimally integrated implies that p_j is increasing, which in turn implies that

$$\frac{p_j}{1-p_j} > \frac{p_{j-1}}{1-p_{j-1}}.$$

This completes the proof. ■

Proof of Proposition 6: I first show that if the player in layer $j \geq 2$ finds it optimal to make no effort, then all players in earlier layers also choose not to make effort. Consider the incentive compatibility condition. Layer j player prefers not

making effort over reducing type-1 error:

$$
\begin{aligned}
c \geq & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \left[\begin{pmatrix} 1-e_1+\Delta & e_1-\Delta \\ 1-e_2 & e_2 \end{pmatrix} - \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2 & e_2 \end{pmatrix} \right] \prod_{n=j+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} \Delta & -\Delta \\ 0 & 0 \end{pmatrix} \prod_{n=j+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \prod_{n=j+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta p_{j-1} \prod_{n=j+1}^{N} (e_{2,n} - e_{1,n}) .
\end{aligned}
\tag{42}
$$

Suppose there is a player in layer $j' < j$ who prefers making effort (either type-1 or type-2) over no effort. Without loss of generality, denote j' to be the first such layer. I claim that the layer-j' player has no incentive to reduce type-1 error. Calculating the payoff difference for this player:

$$
\begin{aligned}
& \begin{pmatrix} p_{j'-1} & 1-p_{j'-1} \end{pmatrix} \left[\begin{pmatrix} 1-e_1+\Delta & e_1-\Delta \\ 1-e_2 & e_2 \end{pmatrix} - \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2 & e_2 \end{pmatrix} \right] \prod_{n=j'+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta p_{j'-1} \prod_{n=j'+1}^{N} (e_{2,n} - e_{1,n}) \\
< & \Delta p_{j-1} \prod_{n=j+1}^{N} (e_{2,n} - e_{1,n}) \leq c.
\end{aligned}
$$

The strict inequality holds because $p_{j'-1} < p_{j-1}$ and $\prod_{n=j'+1}^{j} (e_{2,n} - e_{1,n}) < 1$. The former condition arises from the fact that no one in the initial $j' - 1$ layers makes effort, and condition (27) implies that p_n for $n < j'$ is an increasing sequence. If some players between layer j' and j make effort, the probability p_{j-1} further increases, and $p_{j'-1} < p_j$ remains valid. Hence (42) implies that layer-j' player also does not have incentive to reduce type-1 error.

I next consider effort incentive for reducing type-2 errors. The incentive compat-

ibility condition for layer-j player implies that

$$\begin{aligned} c \geq & \left(\begin{array}{cc} p_{j-1} & 1-p_{j-1} \end{array} \right) \left[\left(\begin{array}{cc} 1-e_1 & e_1 \\ 1-e_2+\Delta & e_2-\Delta \end{array} \right) - \left(\begin{array}{cc} 1-e_1 & e_1 \\ 1-e_2 & e_2 \end{array} \right) \right] \prod_{n=j+1}^{N} \mathcal{M}_n \left(\begin{array}{c} 1 \\ 0 \end{array} \right) \\ = & \left(\begin{array}{cc} p_{j-1} & 1-p_{j-1} \end{array} \right) \left(\begin{array}{cc} 0 & 0 \\ \Delta & -\Delta \end{array} \right) \prod_{n=j+1}^{N} \mathcal{M}_n \left(\begin{array}{c} 1 \\ 0 \end{array} \right) \\ = & \Delta \left(\begin{array}{cc} p_{j-1} & 1-p_{j-1} \end{array} \right) \left(\begin{array}{c} 0 \\ 1 \end{array} \right) \left(\begin{array}{cc} 1 & -1 \end{array} \right) \prod_{n=j+1}^{N} \mathcal{M}_n \left(\begin{array}{c} 1 \\ 0 \end{array} \right) \\ = & \Delta \left(1-p_{j-1}\right) \prod_{n=j+1}^{N} \left(e_{2,n}-e_{1,n}\right). \end{aligned} \tag{43}$$

From the definition of p_{j-1}

$$\left(1-p_{j-1}\right) = \left(1-p_{j-2}\right) e_{2,j-1} - p_{j-2} e_{1,j-1} \geq \left(1-p_{j-2}\right) \left(e_{2,j}-\Delta\right) - p_{j-2} e_{1,j},$$

where the inequality is due to the fact that effort at most reduces $e_{2,j-1}$ by Δ from the status-quo level $e_{2,j} = e_2$, and $e_{1,j-1} \leq e_{1,j} = e_1$. This expression in turn dominates

$$\left(1-p_{j-2}\right) \left(e_{2,j}-\Delta\right) - p_{j-2} e_{1,j} = \left(1-p_{j-2}\right) \left(e_{2,j}-e_{1,j}\right) - \left(1-p_{j-2}\right) \Delta + e_{1,j} > \left(1-p_{j-2}\right) \left(e_{2,j}-e_{1,j}\right),$$

where the last inequality uses the fact that $e_{1,j} = e_1 > \Delta$. Hence,

$$\left(1-p_{j-1}\right) > \left(1-p_{j-2}\right) \left(e_{2,j}-e_{1,j}\right),$$

and condition (43) implies that

$$c > \Delta \left(1-p_{j-2}\right) \prod_{n=j}^{N} \left(e_{2,n}-e_{1,n}\right).$$

Therefore, the player in layer $j-1$ does not have incentive to reduce type-2 error. The same conclusion holds for any $j' \leq j-1$.

Hence, define N_1 to be the last layer that does not make effort. The above proof establishes that all players in layers $n \leq N_1$ do not make effort. Next, I show that if layer-j player prefers reducing type-2 error over type-1 error, then all players in layers $j' < j$ have the same preference. The incentive compatibility condition for layer-j

player implies that

$$\begin{aligned}
0 \geq & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \left[\begin{pmatrix} 1-e_1+\Delta & e_1-\Delta \\ 1-e_2 & e_2 \end{pmatrix} - \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2+\Delta & e_2-\Delta \end{pmatrix} \right] \prod_{n=j+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} \Delta & -\Delta \\ -\Delta & \Delta \end{pmatrix} \prod_{n=j+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \prod_{n=j+1}^{N} \mathcal{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta \left(2p_{j-1}-1\right) \prod_{n=j+1}^{N} \left(e_{2,n}-e_{1,n}\right),
\end{aligned} \tag{44}$$

which is in turn equivalent to $p_{j-1} \leq \frac{1}{2}$.

Suppose in contrast that there is a layer-j' player ($j' < j$) who prefers reducing type-1 error over type-2. Without loss of generality, denote by j' the first layer such that the player prefers reducing type-1 error. An analogous derivation as in (44) yields that $p_{j'-1} \geq \frac{1}{2}$. Therefore, the initial N_1 layers do not make effort and those between N_1+1 and $j'-1$ reduce type-2 error. One can easily calculate the invariant probabilities associated with the status-quo errors $p^*_{(0)} \equiv \frac{1-e_2}{1-e_2+e_1}$, after the reduction of type-1 error $p^*_{(1)} \equiv \frac{1-e_2}{1-e_2+e_1-\Delta}$ and after the reduction of type-2 error $p^*_{(2)} \equiv \frac{1-e_2+\Delta}{1-e_2+\Delta+e_1}$. The supposition that $p_{j'-1} \geq \frac{1}{2}$ implies

$$1-e_2 \geq e_1-\Delta.$$

Otherwise, one can easily verify that $p^*_{(0)}, p^*_{(1)}, p^*_{(2)} \leq \frac{1}{2}$, and no posterior $p_{j'-1} \geq \frac{1}{2}$, creating a contradiction. Hence, the ranking of the three invariant probabilities must be

$$p^*_{(0)} < p^*_{(2)} \leq p^*_{(1)}. \tag{45}$$

To verify, note that

$$\begin{aligned}
& p^*_{(1)} \geq p^*_{(2)} \\
\Leftrightarrow \quad & \tfrac{1-e_2+\Delta+e_1}{1-e_2+e_1-\Delta} \geq \tfrac{1-e_2+\Delta}{1-e_2} \\
\Leftrightarrow \quad & \tfrac{2\Delta}{1-e_2+e_1-\Delta} \geq \tfrac{\Delta}{1-e_2} \\
\Leftrightarrow \quad & 2\left(1-e_2\right) \geq 1-e_2+e_1-\Delta \\
\Leftrightarrow \quad & 1-e_2 \geq e_1-\Delta.
\end{aligned}$$

However, the ranking in (45) still leads to a contradiction. First, it must be that $p_{j'-1} < p^*_{(2)}$ because no one prior to layer j' reduces type-1 error. Furthermore,

all players between layers j' and j make either type-1 or type-2 efforts, hence the posterior $p_{j-1} > p_{j'-1} \geq \frac{1}{2}$. A contradiction.

Hence, let N_2 be the last layer that reduces type-2 error, then all layers between N_1+1 and N_2 make effort to reduce type-2 error, and all layers after N_2 reduce type-1 error. This completes the proof. ■

Proof of Proposition 7: As explained in the main text, the optimal integration rule reduces to comparing $\mathbf{P}_0 (\mathcal{M}_H - \mathcal{I}) \mathcal{M}_{AI} \mathbf{U}'$ and $\mathbf{P}_0 \mathcal{M}_{AI} (\mathcal{M}_H - \mathcal{I}) \mathbf{U}'$.

First, consider human in the first layer: $\mathbf{P}_0 (\mathcal{M}_H - \mathcal{I}) \mathcal{M}_{AI} \mathbf{U}'$. Simple calculation yields

$$\begin{aligned} \mathbf{P}_0 (\mathcal{M}_H - \mathcal{I}) &= \begin{pmatrix} 0 & p_0 & 1-p_0 \end{pmatrix} \begin{pmatrix} -e_{1,H} & 0 & e_{1,H} \\ 0 & -e_{1,H} & e_{1,H} \\ 0 & 1-e_{2,H} & e_{2,H}-1 \end{pmatrix} \\ &= [-p_0 e_{1,H} + (1-p_0)(1-e_{2,H})] \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}, \end{aligned} \tag{46}$$

and

$$\begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} a & b & c \\ s_{AI} & 1-s_{AI}-e_{1,AI} & e_{1,AI} \\ s_{AI} & 1-s_{AI}-e_{2,AI} & e_{2,AI} \end{pmatrix} = (e_{2,AI} - e_{1,AI}) \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}.$$

Hence,

$$\mathbf{P}_0 (\mathcal{M}_H - \mathcal{I}) \mathcal{M}_{AI} \mathbf{U}' = [-p_0 e_{1,H} + (1-p_0)(1-e_{2,H})] (e_{2,AI} - e_{1,AI}). \tag{47}$$

Next, consider AI in the first layer: $\mathbf{P}_0 \mathcal{M}_{AI} (\mathcal{M}_H - \mathcal{I}) \mathbf{U}'$. Simple calculation yields

$$\begin{aligned} \mathbf{P}_0 \mathcal{M}_{AI} &= \begin{pmatrix} 0 & p_0 & 1-p_0 \end{pmatrix} \begin{pmatrix} a & b & c \\ s_{AI} & 1-s_{AI}-e_{1,AI} & e_{1,AI} \\ s_{AI} & 1-s_{AI}-e_{2,AI} & e_{2,AI} \end{pmatrix} \\ &= \begin{pmatrix} s_{AI} & 1-s_{AI}-p_0 e_{1,AI} - (1-p_0) e_{2,AI} & p_0 e_{1,AI} + (1-p_0) e_{2,AI} \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned}
& \mathbf{P}_0 \mathcal{M}_{AI} \left(\mathcal{M}_H - \mathcal{I} \right) \mathbf{U}' \\
= & \left(\begin{array}{ccc} s_{AI} & 1 - s_{AI} - p_0 e_{1,AI} - (1-p_0) e_{2,AI} & p_0 e_{1,AI} + (1-p_0) e_{2,AI} \end{array} \right) \left(\begin{array}{c} -S e_{1,H} \\ -e_{1,H} \\ 1 - e_{2,H} \end{array} \right) \\
= & -e_{1,H} \left[s_{AI} (S-1) + 1 - p_0 e_{1,AI} - (1-p_0) e_{2,AI} \right] + (1 - e_{2,H}) \left[p_0 e_{1,AI} + (1-p_0) e_{2,AI} \right].
\end{aligned}$$

Comparing the difference

$$\begin{aligned}
& \mathbf{P}_0 \left(\mathcal{M}_H - \mathcal{I} \right) \mathcal{M}_{AI} \mathbf{U}' - \mathbf{P}_0 \mathcal{M}_{AI} \left(\mathcal{M}_H - \mathcal{I} \right) \mathbf{U}' \\
= & \left[-p_0 e_{1,H} + (1-p_0)(1 - e_{2,H}) \right] (e_{2,AI} - e_{1,AI}) \\
& + e_{1,H} \left[s_{AI} (S-1) + 1 - p_0 e_{1,AI} - (1-p_0) e_{2,AI} \right] - (1 - e_{2,H}) \left[p_0 e_{1,AI} + (1-p_0) e_{2,AI} \right] \\
= & -p_0 e_{1,H} (e_{2,AI} - e_{1,AI}) + (1-p_0)(1 - e_{2,H})(e_{2,AI} - e_{1,AI}) \\
& + e_{1,H} \left[s_{AI} (S-1) + 1 - p_0 e_{1,AI} - (1-p_0) e_{2,AI} \right] - (1 - e_{2,H}) \left[p_0 e_{1,AI} + (1-p_0) e_{2,AI} \right] \\
= & e_{1,H} \left[s_{AI} (S-1) + 1 - e_{2,AI} \right] - (1 - e_{2,H}) e_{1,AI}
\end{aligned}$$

Therefore, human should be the first layer iff

$$\frac{e_{1,H}}{1 - e_{2,H}} \left[\frac{s_{AI}}{1 - e_{2,AI}} (S-1) + 1 \right] - \frac{e_{1,AI}}{1 - e_{2,AI}} \geq 0,$$

and hence condition (34).

It is clear from (32) that $\mathbf{P}_0 \left(\mathcal{M}_H - \mathcal{I} \right) \mathcal{M}_{AI} \mathbf{U}' \geq \mathbf{P}_0 \mathcal{M}_{AI} \left(\mathcal{M}_H - \mathcal{I} \right) \mathbf{U}'$ immediately implies $\pi^*_{H,1} \geq \pi^*_{H,2}$.

Finally, comparing $\mathbf{P}_0 \left(\mathcal{M}_H - \mathcal{I} \right) \mathbf{U}'$ and $\mathbf{P}_0 \left(\mathcal{M}_H - \mathcal{I} \right) \mathcal{M}_{AI} \mathbf{U}'$, it is clear from (46) and (47) that the former is greater. Following the same logic, (34) and (33) immediately imply that $\pi^*_{H,1} \leq \pi^*_{H,0}$. ■

Proof of Proposition 8: Denote by I the $(N+1) \times (N+1)$ dimensional identity matrix; by I_n the same dimensional matrix with ones in the nth column and zeros everywhere else. Therefore,

$$\mathcal{M}_n = (1 - q_n - q'_n) I + q_n I_n + q'_n I_{N+1}.$$

Before proceeding, it is useful to note that $I_n I_m = I_m$ for any $n, m \leq N+1$. Therefore,

$$(I_n - I)\, I_m = 0. \tag{48}$$

First, suppose the agent in layer $i = \sigma(n) > 0$ is the first layer in the delegation sequence with $U_n < 0$. Eliminating this agent from the delegation process is equivalent to replacing $\mathcal{M}_n$ by I in payoff calculation (5). Calculating the difference:

$$\begin{array}{rl} & \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \left(\mathcal{M}_n - I\right) \prod_{j>i} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}' \\ = & \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \left[q_n \left(I_n - I\right) + q'_n \left(I_{N+1} - I\right)\right] \prod_{j>i} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}' \\ = & \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \left[q_n \left(I_n - I\right) + q'_n \left(I_{N+1} - I\right)\right] \left[y_0 I + \sum_{j=i+1}^{I} y_j I_{\sigma^{-1}(j)} + y_N I_N\right] \mathbf{U}' \end{array},$$

where y_j are some coefficients. Using condition (48), the above difference becomes

$$\begin{array}{rl} & y_0 \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \left[q_n \left(I_n - I\right) + q'_n \left(I_{N+1} - I\right)\right] \mathbf{U}' \\ = & y_0 q_n \left(U_n - \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}'\right) + y_0 q'_n \left(U_{N+1} - \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}'\right), \\ < & -y_0 \left(q_n + q'_n\right) \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}' \end{array}$$

where the last inequality utilizes the fact that $U_n < 0$ and $U_{N+1} = 0$. Since the initial $i-1$ agents are associated with positive U_j's, the last expression is negative. Hence, removing agent n strictly improves the expected payoff to the principal.

Next, suppose there are two adjacent agents $i = \sigma^{-1}(n_1)$ and $i+1 = \sigma^{-1}(n_2)$ such that

$$\frac{U_{n_1}}{1 + \frac{q'_{n_1}}{q_{n_1}}} > \frac{U_{n_2}}{1 + \frac{q'_{n_2}}{q_{n_2}}}.$$

Consider the payoff difference with their positions reversed:

$$\mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \left(\mathcal{M}_{n_1} \mathcal{M}_{n_2} - \mathcal{M}_{n_2} \mathcal{M}_{n_1}\right) \prod_{j>i+1} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}'.$$

Calculate

$$\begin{array}{rl} & \mathcal{M}_{n_1} \mathcal{M}_{n_2} \\ = & \left(1 - q_{n_1} - q'_{n_1}\right)\left(1 - q_{n_2} - q'_{n_2}\right) I + q_{n_1}\left(1 - q_{n_2} - q'_{n_2}\right) I_{n_1} + q_{n_2} I_{n_2}, \\ & + \left[q'_{n_1}\left(1 - q_{n_2} - q'_{n_2}\right) + q'_{n_2}\right] I_{N+1} \end{array}$$

and using symmetry to take the difference

$$\begin{array}{rl} & \mathcal{M}_{n_1}\mathcal{M}_{n_2} - \mathcal{M}_{n_2}\mathcal{M}_{n_1} \\ = & -q_{n_1}\left(q_{n_2}+q'_{n_2}\right) I_{n_1} + q_{n_2}\left(q_{n_1}+q'_{n_1}\right) I_{n_2} + xI_{N+1} \end{array},$$

where x is some constant.

Therefore, the payoff difference becomes

$$\begin{array}{rl} & \mathbf{P}_0 \prod_{j<i} \mathcal{M}_{\sigma^{-1}(j)} \left(\mathcal{M}_{n_1}\mathcal{M}_{n_2} - \mathcal{M}_{n_2}\mathcal{M}_{n_1}\right) \prod_{j>i+1} \mathcal{M}_{\sigma^{-1}(j)} \mathbf{U}' \\ = & -q_{n_1}\left(q_{n_2}+q'_{n_2}\right) U_{n_1} + q_{n_2}\left(q_{n_1}+q'_{n_1}\right) U_{n_2} \\ = & -\left(q_{n_1}+q'_{n_1}\right)\left(q_{n_2}+q'_{n_2}\right)\left(\frac{U_{n_1}}{1+\frac{q'_{n_1}}{q_{n_1}}} - \frac{U_{n_2}}{1+\frac{q'_{n_2}}{q_{n_2}}}\right) < 0 \end{array},$$

a contradiction to optimality and concluding the proof. ■