Efficient Integration: Human, Machine, and Generative AI

Hongda Zhong, University of Texas at Dallas and CEPR

June 23, 2024

Abstract

I study the optimal integration of humans and technologies in multi-layered decision-making processes. Each layer corrects errors made by previous layers but also introduces new errors. The ratio between these two aspects serves as a one-dimensional quality measure that uniformly ranks all technologies. Optimal integration requires higher quality technologies to be applied later. Furthermore, technologies asymmetrically influence human effort. Initial players specialize in correcting errors, while final players focus on avoiding new errors. When error costs depend on unobservable states, a uniform ranking is not generally possible. Consequently, the introduction of new technologies can significantly disrupt the integration of existing ones.

Key Words: Multi-layered decision making, Automation, Credit screening, Academic promotion, Error reduction

1 Introduction

Modern production and decision-making processes increasingly rely on automation, with recent years witnessing the rapid proliferation of generative artificial intelligence (AI) technologies such as ChatGPT and Full Self-Driving. While these technologies sometimes outperform human decisions, they also introduce new errors into the system. Several fundamental questions naturally arise: How should humans and technologies be integrated efficiently? Who should be the ultimate decision-maker? How does this integration influence human effort? How does generative AI differ from traditional technologies? Why does it create such a profound impact?

In this paper, I propose a framework based on linear algebra to analyze these questions. Specifically, I examine a sequential decision-making process. Each layer of decision technology (human or machine) takes the proposed action from the preceding layer as input and recommends a (possibly random) action for the subsequent layer. The final layer's recommendation becomes the ultimate action. Each action generates a payoff that may depend on some unobservable states. The optimal integration of technologies (and humans) is the particular order that maximizes the expected payoff.

The key object in the decision-making process is the distribution of the proposed actions after each layer. As the technology in each layer alters the proposed actions, the effect is mathematically characterized by the corresponding probability transition matrix on the state-action space. The optimal integration is, therefore, represented by a particular sequence of these transition matrices that maximizes the final distribution's payoff. It is worth noting that there is no easy criterion to generally determine whether two matrices commute, making the problem of finding the optimal sequence for multiplying a set of matrices more challenging. For tractability and interpretation, this paper predominantly focuses on the case of binary actions, which arguably covers numerous applications.

For instance, in the aviation or automobile industries, autopilot or self-driving functions increasingly take over controls from humans. With a great level of abstraction, one can broadly categorize actions into safe or dangerous ones. Both humans and automation can make two types of mistakes: incorrectly altering a safe action to a dangerous action (type-1 error, denoted by e_1) and failing to correct a dangerous action into a safe one (type-2 error, denoted by e_2). The probability transition matrices in this case are of size 2×2, determined by the chances of making the two types of errors.

The analysis of optimal integration sheds light on whether automation should be given the authority to override human inputs in some scenarios.

In this simplest case with binary actions, the invariant probability ($\frac{1-e_2}{1-e_2+e_1}$) associated with the 2×2 transition matrix serves as a one-dimensional "quality measure" that uniformly ranks all decision technologies. A technology shifts the prior probability of a safe action through the transition matrix, steering it towards its invariant probability. The ranking by invariant distribution is also equivalent to the ranking by the ratio between error correction $(1 - e_2)$ and the new errors (e_1) introduced by the technology.

Efficient integration follows a simple rule: Apply technologies in ascending order of their quality (equivalently, their invariant probabilities). Intuitively, efficient integration should feature "superior" technologies serving as later-stage "gatekeepers" in the decision-making process. Otherwise, a safe action proposed by a superior technology early on may be erroneously altered by a subsequent, inferior technology, leading to worse outcomes. Technologies with invariant probabilities lower than the prior also degrade the outcome and therefore should be eliminated. Under efficient integration, the likelihood of achieving a correct (safe) action increases progressively throughout the decision-making process.

Type-1 error can be interpreted as execution quality (the ability to maintain safe actions) whereas type-2 error measures troubleshooting capabilities (the ability to correct dangerous actions). Traditional technologies are good at execution ($e_1 \approx 0$), but can rarely troubleshoot ($e_2 \approx 100\%$). The invariant probability becomes "$\frac{0}{0}$" indeterminate form. The optimal placement of traditional technologies is therefore very sensitive to the specific combination of the two errors, which explains why similar automotive technologies may be placed differently by different companies. For example, 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 will override those actions. In contrast, Boeing maintained the alternative design philosophy that humans should have ultimate control for many years (Kornecki and Hall 2004).

Moreover, the model offers insights into the significance of generative AI, such as ChatGPT or Full Self-Driving. Relative to traditional technologies, generative AI aligns more closely with human capabilities, occasionally demonstrating troubleshooting abilities that reduce type-2 errors from approaching 100%. For instance, Full Self-Driving technology can autonomously avert collisions without driver input. Consequently, the invariant probability of generative AI rapidly approaches 100%. The efficient integration rule therefore calls for AI to be the ultimate decision-maker. It's important to note that this structure does not exclude humans. Instead, humans initially focus on initial decision-making and troubleshooting, followed by final execution through AI.

In addition, this outcome does not rely on AI's type-2 error being lower than that of humans. AI provides an additional avenue for error correction without introducing many new errors, thanks to its execution quality (low type-1 error). The key to successful AI lies in its marginal reduction in type-2 errors while maintaining a low type-1 error. Therefore, AI hallucination that compromises the executional quality of the AI poses a significant threat to the technology's applicability.

I then investigate how human efforts to mitigate errors are influenced by their placement within the integration. Intriguingly, the incentive for effort across layers reveals distinct patterns concerning the reduction of type-1 and type-2 errors. Human effort incentives are contingent upon two factors: the relevance of a particular type of error faced by humans and the consequence of human errors on the probability of achieving the correct final action.

The consequences of errors, regardless of type, are more significant in later stages due to fewer opportunities for subsequent layers to rectify such errors. However, the relevance of the two types of errors is asymmetrical. Type-1 errors, stemming from correct inputs, become increasingly relevant in later stages of the process, where correct actions are more likely to be proposed. Conversely, type-2 errors, stemming from incorrect inputs, diminish in relevance as the likelihood of erroneous inputs decreases. Therefore, the incentive for effort to mitigate type-1 errors increases as the process progresses, driven by both the increasing consequence and relevance. Similarly, effort incentives for type-2 errors increase, albeit at a slower pace, as the greater consequence outweighs the diminishing relevance.

This observation explains the phenomenon where the presence of technologies capable of assuming ultimate control often dampens human effort incentives. For instance, human pilots or drivers may be less vigilant when relying on autopilot systems for aircraft or vehicle operation. On the positive side, these technologies also free humans from executional tasks, allowing them to focus more on troubleshooting efforts in the case of emergencies.

This insight into effort patterns also extends to scenarios involving multiple layers of human players. Players who are ex-ante identical endogenously specialize in mitigating different types of errors based on their position within the process. Initial players exhibit minimal effort in error reduction, as their errors carry negligible consequences, given the presence of subsequent players capable of potentially rectifying errors. Players positioned in middle layers specialize in mitigating type-2 errors, as input actions are still noisy, rendering type-2 errors more relevant. Finally, players situated at the conclusion of the process concentrate on reducing type-1 errors, as input actions are more likely to be correct following numerous layers of selection.

A pertinent example is the academic promotion (tenure) process, which frequently involves multiple layers of evaluation, including assessments from letter writers, departmental committees, schools, the provost's office, and final decisions from university presidents. Initial layers often entail substantial effort devoted to screening promotion cases for merit, thereby rectifying errors and addressing misjudgments. Conversely, at the level of provosts or presidents, decision-makers seldom overturn department or school recommendations. Instead, they focus on process execution and ensure adherence to correct procedures.

Finally, I consider multiple fundamental states, where errors stemming from different underlying types may incur varying costs. For instance, in loan screening, there are two unobservable states: good or bad borrowers, and two actions for the bank: to lend or to reject the applicant. Lending to a risky borrower may incur significant costs due to potential loss of principal, whereas rejecting a creditworthy borrower may only result in a small loss from potential interest income. The payoff of a decision is defined on the four-element (2×2) state-action space: accepting/rejecting good/bad applicants. A generic technology is mathematically characterized by a 4×4 transition matrix on the state-action space.

The analysis of this case is significantly more involved and generates new insights. First, when there are only two available technologies, I provide an explicit condition for determining which technology should be applied first. This condition incorporates expected cost of errors and the quality of the technology conditional on each state. However, surprisingly, this binary relation between technologies is not transitive when more than two technologies are available. Specifically, it is possible to have three technologies $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ forming a circular order, such that $\mathcal{M}_1\mathcal{M}_2$, $\mathcal{M}_2\mathcal{M}_3$, and $\mathcal{M}_3\mathcal{M}_1$ are the optimal order when only each respective pair is available.

The lack of transitivity delivers an interesting insight: The introduction of a new technology may significantly alter how existing decision-making processes are integrated. In the previous example, suppose the most efficient integration of all three technologies is in the order $\mathcal{M}_3\mathcal{M}_1\mathcal{M}_2$. This implies that the introduction of $\mathcal{M}_1$ changes the relative position of $\mathcal{M}_2$ and $\mathcal{M}_3$ in the efficient integration. In fact, given the circular binary order, any three-technology integration will result in a reversal of positions for two of the technologies. This feature provides another reason why a new technology such as generative AI can be revolutionary for existing decision-making processes.

This paper contributes to several strands of literature in decision science and economics. Within the vast field of decision science, one particularly relevant approach is the widely accepted Analytic Hierarchy Process (AHP) in Multiple-Criteria Decision Making (MCDM), pioneered by Saaty (1977). Utilizing linear algebra, AHP provides a robust method for determining the best option among many candidates, each with multi-dimensional attributes. My work complements this method by solving for the optimal sequence of applying different decision-making technologies. While both analyses utilize linear algebra, the problems addressed are distinct. AHP employs eigenvectors to extract the relative priority of attributes from pairwise preferences, whereas the core mathematical problem in my analysis is determining the optimal sequence for multiplying matrices when they generally do not commute.

The methodology developed in this paper relates to Zhong (2023), where a special case of this framework is used to analyze the distribution of skill and effort along intermediation chains. However, the current framework is significantly more general, incorporating multiple states and actions and solving for the optimal sequence of different layers.

Regarding the applications analyzed, the paper investigates the effort problem involving multiple economic agents. Holmstrom (1982) studies moral hazard problems in teams, but his analysis lacks the feature of hierarchical structure. Additionally, this paper introduces the novel separation of two types of effort: correcting mistakes and improving execution, demonstrating how these efforts evolve differently along the decision-making chain.

Two strands of literature in economics consider the multi-layered structure: hierarchy in firms (Calvo and Wellisz 1979; Qian 1994; Chen 2017; Garicano 2000) and intermediation chains often observed in financial markets (Glode and Opp 2016; Glode, Opp, and Zhang 2019; He and Li 2023; Dasgupta and Maug 2021). This paper contributes to these literatures by introducing the concept of optimal sequencing of different technologies and applying it in the context of artificial intelligence.

2 Baseline Model Setup

2.1 Model

Consider a sequential decision-making process where there are T unobservable types of fundamental states and A different available actions. Let $\mathbb{T}$ and $\mathbb{A}$ denote the sets of types and actions, respectively. The probability distribution of fundamental types is represented by $\mathbf{q} = (q_1, q_2, ..., q_T)$. Action a taken under fundamental type t creates a certain payoff (or utility) U_{at}, and these payoffs are collectively captured by a real vector

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

The final outcome of the decision-making process can be characterized by a probability distribution of actions $\mathbf{p}_t = (p_{1t}, p_{2t}, ..., p_{At})$, which may depend on the fundamental type t. Together with the fundamental distribution $\mathbf{q}$, the decision induces a probability distribution on the AT dimensional state-action space

$$\mathbf{P} = (q_1p_{11}, q_1p_{21}, ..., q_1p_{A1}, q_2p_{12}, q_2p_{22}, ..., q_2p_{A2}, ..., q_Tp_{1T}, ..., q_Tp_{AT}). \quad (2)$$

Therefore, the expected payoff is given by $\mathbf{P}\mathbf{U}^{'}$, where the prime in the superscript denotes transpose.

There are $N \geq 2$ decision-making technologies (or human players) that can be integrated sequentially. Each technology takes a recommended action $a_{i-1} \in \mathbb{A}$ from the previous layer as input and produces an output action a_i for the next layer. Technologies may create randomness, where the output action follows a distribution on $\mathbb{A}$ depending on the fundamental type. The initial input action a_0 to the first technology is assumed to follow an exogenous prior distribution $\mathbf{P}_0$. The action recommended by the final layer is adopted as the ultimate action.

A generic technology $\mathcal{M}$ is mathematically characterized by an $AT \times AT$ dimensional probability transition matrix with T $A \times A$-sized diagonal blocks $\mathcal{M}^{(t)}$

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}^{(1)} & 0 & 0 & 0 \\ 0 & \mathcal{M}^{(2)} & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \mathcal{M}^{(T)} \end{pmatrix}. \tag{3}$$

After applying the technology, the posterior probability distribution of actions is given by $\mathbf{P}_1 = \mathbf{P}_0\mathcal{M}$. To clarify, a technology can alter the proposed action, but not the fundamental state.[1] Thus, all off-diagonal blocks are zero, indicating no transition between fundamental types. The diagonal blocks describe how the technology modifies proposed actions and can accommodate stochastic outcomes. Specifically, if the input action is a, then the output action follows a probability distribution on $\mathbf{A}$ according to the a-th row in matrix $\mathcal{M}^{(t)}$. The type-dependent matrices $\mathcal{M}^{(t)}$ also allow for arbitrary learning of the underlying fundamental state in the decision-making process.

When N technologies $\{\mathcal{M}_n\}$ are applied sequentially, the posterior probability distribution of actions is given by

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

and the payoff from the decision is

$$\mathbf{P}_N \mathbf{U}' = \mathbf{P}_0 \prod_{n=1}^{N} \mathcal{M}_n \mathbf{U}'. \tag{5}$$

Efficient integration for a given prior $\mathbf{P}_0$ and payoff vector $\mathbf{U}$ is a permutation of a subset of technologies that maximizes the payoff. Specifically, integration is represented by a mapping $\sigma : \{1, 2, ..., N\} \to \{0, 1, 2, ..., I\}$ for some $I \le N$ such that $\sigma^{-1}(i)$ uniquely exists for each $i = 1, 2, ..., I$. In this mapping, only I out of N technologies are included in the integration, and $\sigma(n)$ denotes the new position of the original technology n within the integration. A technology is excluded if $\sigma(n) = 0$.

The goal is to find the most efficient integration $\sigma^*(\cdot)$ that maximizes the payoff from the final

[1] While not crucial for the insight, this is a natural restriction for all applications in Section 2.2. For example, when bank makes lending decision, applicants' intrinsic creditworthiness should not be affected by banks' screening technologies.

decision

$$\max_{\sigma(\cdot),I} \mathbf{P}_0 \prod_{i=1}^{I} \mathcal{M}_{\sigma^{-1}(i)} \mathbf{U}'. \tag{6}$$

2.2 Applications of the Model

2.2.1 Automation – Autopilot and Self-Driving

The simplest case of the general model, $T = 1$ and $A = 2$, reflects automation in production, such as in the transportation industry (aviation or automobile). Using the example of aviation, there is a single underlying fundamental type ($T = 1$)—safely transporting passengers to their destinations—and two possible actions: safe or disastrous ($A = 2$). If a safe action is chosen, the normalized payoff is 1; otherwise, for a disastrous action, the payoff is 0. Hence, the payoff vector is $\mathbf{U} = (1, 0)$.

It's widely recognized that human pilots are susceptible to errors, which can lead to disasters. To mitigate this risk, a second pilot (copilot) is introduced into the cockpit, tasked with cross-checking each other's actions. However, this system is not infallible, as the copilot may incorrectly alter a correct action taken by the pilot. With technological advancements, autopilot systems have been introduced to handle routine controls previously managed by human pilots. Yet, this solution is also not without flaws; technology failures can occur, and automation might also make human pilots complacent. The model effectively captures these complexities.

A general decision maker, whether human pilots or autopilot, can be represented by the following 2×2 matrix

$$\mathcal{M}_n = \begin{pmatrix} 1 - e_{1,n} & e_{1,n} \\ 1 - e_{2,n} & e_{2,n} \end{pmatrix}, \tag{7}$$

where $e_{1,n}$ denotes the probability of a type-1 error – incorrectly changing a safe action to a disastrous action and $e_{2,n}$ denotes the probability of a type-2 error – failure 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 assumption that type-2 error is typically higher than type-1 error ($e_2 > e_1$) is well-suited in aviation contexts. Maintaining routine cruising during normal flights often involves minimal errors and few things can go wrong (lower type-1 error). In contrast, recovering from unusual flight attitudes is considerably more challenging (higher type-2 error).

The model's solution provides insights into how automation should be integrated into the decision-making process most effectively. It addresses questions like whether human pilots should retain ultimate authority or if autopilot should have final control under certain circumstances.

2.2.2 Loan Screening

Introducing uncertainty in fundamental states $T \geq 2$, the model can be applied to a bank's loan screening problem. Consider $T = 2$ types of borrowers: (G)ood or (B)ad, and $A = 2$ different actions – (A)ccept or (R)eject. The ratio of good borrowers is π_G, and thus the ratio of bad borrowers is $\pi_B = 1 - \pi_G$.

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, regardless of type, yields a payoff of 0. Therefore, the payoff vector is $\mathbf{U} = (r, 0, 0, -L)$.

As a special case of the general specification (3), a loan screening technology (whether a human loan officer or an automated processing system) can be described by a probability transition matrix considering the four possible outcomes: $\{GA, GR, BR, BA\}$,

$$\mathcal{M}_n = \begin{pmatrix} 1 - e_{1,n}^{(G)} & e_{1,n}^{(G)} & 0 & 0 \\ 1 - e_{2,n}^{(G)} & e_{2,n}^{(G)} & 0 & 0 \\ 0 & 0 & 1 - e_{1,n}^{(B)} & e_{1,n}^{(B)} \\ 0 & 0 & 1 - e_{2,n}^{(B)} & e_{2,n}^{(B)} \end{pmatrix}. \tag{8}$$

As in (3), it features $T = 2$ diagonal blocks, each of size 2×2 ($A \times A$), capturing the type-1 and type-2 errors associated with good and bad borrowers: $e_{1,n}^{(\theta)}$ and $e_{2,n}^{(\theta)}$ ($\theta = G$ or B). To elaborate, for a good borrower, the "correct" action—yielding the higher payoff—is to accept, whereas for a bad borrower, the correct action is to reject. Therefore:

- For good borrowers, a type-1 error is to switch from accepting (GA) to rejecting (GR) and a type-2 error is to maintain the GB state.

- For bad borrowers, a type-1 error is to switch from rejecting (BR) to accepting (BA) and a type-2 error is to maintain the BA state.

Technologies can change lending decisions, but do not alter the borrower type, which results in the block diagonal form of $\mathcal{M}_n$.

2.2.3 Multi-Layered Decision Makers – Academic Promotion

Another enhancement involves endogenizing the technologies in each layer $\mathcal{M}_n(\mathbf{f}_n)$ by linking them to the effort choice $\mathbf{f}_n$ of each decision maker. This approach is particularly relevant when analyzing multi-layered human decision-making processes rather than relying on exogenous technologies like autopilots.

For example, a typical promotion process in academia includes evaluations and recommendations from external letter writers, a departmental committee, a school committee, the provost's office, and ultimately the president's final decision. The model, enriched with effort choices as discussed in Subsection 4.2 offers insights into the specific aspects each committee should prioritize. Moreover, the model can illustrate how automation (such as autopilots or self-driving systems) influences the effort exerted by human decision makers.

3 Efficient Integration, Solution with $T = 1$ and $A = 2$

I start with the simplest case with binary actions ($A = 2$) and no fundamental uncertainty ($T = 1$), motivated by the application in automation as discussed in Section 2.2.1. This analysis delivers several key insights that serve as the foundation for more general cases and enriched models.

Throughout this section, I use the simplified notation from equation (7) and specifically consider the more natural scenario where $e_{2,n} > e_{1,n}$: Correcting a wrong action is more challenging than maintaining a correct one. The payoff vector is normalized to $\mathbf{U} = (1, 0)$ without loss of generality.

To determine the optimal integration strategy, I first characterize the impact of applying a technology $\mathcal{M}_i$ on the success probability. Define the invariant probability p_i^* associated with $\mathcal{M}_i$ as

$$\begin{pmatrix} p_i^* & 1 - p_i^* \end{pmatrix} = \begin{pmatrix} p_i^* & 1 - p_i^* \end{pmatrix} \mathcal{M}_i. \tag{9}$$

A simple manipulation yields

$$p_i^* = \frac{1 - e_{2,i}}{1 - e_{2,i} + e_{1,i}}. \tag{10}$$

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

The evolution of probability, as expressed in equation (4) immediately implies $\mathbf{P}_i \equiv (p_i, 1 - p_i) = \mathbf{P}_{i-1}\mathcal{M}_i$, or explicitly,

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

where p_i denotes the probability of the correct action being proposed after i layers of decision making.

Taking the difference between (9) and (11) yields

$$p_i^* - p_i = (e_{2,i} - e_{1,i})\,(p_i^* - p_{i-1})\,, \tag{12}$$

or equivalently,

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

This indicates that a technology $\mathcal{M}_i$ shifts the prior probability towards its invariant probability p_i^*. The difference between these probabilities diminishes by a factor of $e_{2,i} - e_{1,i}$, and the posterior probability p_i lies between the prior probability p_{i-1} and the invariant probability p_i^*. With this intuition, we are prepared to characterize efficient integration strategies.

Lemma 1 ***[Ranking of Technologies]*** *Consider two technologies $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $p_0 \le p_1^* \le p_2^*$. Then, applying $\mathcal{M}_1$ first achieves a weakly higher payoff than applying $\mathcal{M}_2$ first. Formally,*

$$\begin{pmatrix} p_0 & 1 - p_0 \end{pmatrix} \mathcal{M}_1\mathcal{M}_2\mathbf{U}' \ge \begin{pmatrix} p_0 & 1 - p_0 \end{pmatrix} \mathcal{M}_2\mathcal{M}_1\mathbf{U}', \tag{14}$$

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

Lemma 1 is a powerful result, enabling the ranking of technologies based on two critical dimensions: type-1 and type-2 errors. Intuitively, to maximize the eventual probability of correctness, it is more effective to position the superior technology (with higher p_i^*) later in the sequence as the "gatekeeper." Conversely, placing a superior technology early is ineffective if subsequent technologies introduce more errors and compromise correct actions. However, comparing two technologies directly can be challenging due to multi-dimensional error profiles: one technology may create higher type-1 errors while the other generates higher type-2 errors. Lemma 1 reduces the dimensionality of the problem and shows that the invariant probability is the key quantity to rank different technologies.

The binary relationship between two technologies given by Lemma 1 is transitive. I can therefore extend the intuition from Lemma 1 to characterize the efficient integration for an arbitrary number of technologies. It's important to note that this transitivity does not hold in more general cases, such as when $T \geq 2$. Detailed analysis of the general and more complex cases is deferred to Section 5.

Proposition 1 ***[Efficient 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 efficient 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 $p_{n_1}^* < p_{n_2}^*$.

Proposition 1 provides a concise and powerful rule for determining the most efficient sequential integration of multiple technologies. To implement this, one simply calculates the invariant probabilities p_n^* associated with each of the N technologies, eliminates those with $p_n^* < p_0$, and then applies the remaining technologies sequentially from lowest to highest invariant probabilities.

Intuitively, because a technology shifts the prior probability p_0 towards its invariant probability p^*, applying a technology with $p^* < p_0$ worsens the posterior relative to the prior. Therefore, an efficient integration strategy excludes such technologies. Conversely, technologies with $p^* > p_0$ strictly improve the outcome and should thus be included in the integration. Additionally, the insight from Lemma 1 that superior technologies (those with higher p_i^*) should be positioned later in the sequence extends naturally to the case $N > 2$.

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 efficient integration implies that the sequence of p_i is also increasing. This will be particularly relevant when considering human effort in decision-making processes.

Corollary 1 *Under the efficient integration of technologies, the probability p_i increases in i.*

Implications for Automation

Revisiting the application discussed in Section 2.2.1, the baseline model provides insights into several fundamental questions regarding technology and generative AI. When should humans have the final authority, and when should technology have the final authority? What sets generative AI apart from traditional technologies, and why does it have such a profound impact on our daily lives?

Compared to humans, technology excels in execution capabilities. For example, cruise control in cars can maintain a precise speed (e.g., 60 mph) with minimal deviations, often better than a human driver. Similarly, autopilot systems in airplanes can navigate a set course with greater precision than human pilots. The ability to execute correct inputs and maintain accurate outputs is measured by the type-1 error e_1 (or more precisely, $1 - e_1$) in the model. Machines rarely generate a type-1 error and fail to execute a proper input. For example, in autopilot systems, failure to execute may occur if a critical sensor fails. Mathematically, $e_{1,m} = \epsilon \to 0$. In contrast, humans typically have a larger type-1 error $e_{1,h} > \epsilon$.

Conversely, humans excel in troubleshooting or innovation capabilities, meaning they can deliver correct actions even when the initial input is incorrect. For instance, a human driver might steer away from a potential crash, whereas traditional cruise control lacks this adaptive capability. In aviation emergencies, human pilots can often navigate situations where traditional autopilot systems cannot handle. The ability to troubleshoot, or correct mistakes, is captured by the type-2 error e_2 (or more precisely, $1 - e_2$) in the model. Machines with limited adaptive capabilities may have a high type-2 error $e_{2,m} = 1 - \delta \to 100\%$, whereas humans generally exhibit a lower type-2 error $e_{2,h} < 1 - \delta$.

Given these error patterns of humans ($e_{i,h}$, $i = 1, 2$) and machine ($e_{i,m}$), Proposition 1 offers implications on the utility of machines and the allocation of final authority between humans and machines.

Corollary 2 *A machine should be integrated into the decision-making process if* $\frac{\delta}{\delta+\epsilon} > p_0$*. Assuming the human is part of the efficient integration, i.e.* $\frac{1-e_{2,h}}{1-e_{2,h}+e_{1,h}} > p_0$*, a machine should monitor human decisions and have final authority if*

$$\frac{\delta}{\delta + \epsilon} > \frac{1 - e_{2,h}}{1 - e_{2,h} + e_{1,h}}, \tag{15}$$

Otherwise, the human should monitor the machine's decisions and have final authority.

An interesting observation is that the usefulness of a technology and its integration into production are highly sensitive to how close its type-1 error $e_{1,m}$ is to 0 and its type 2 error $e_{2,m}$ is to 100%. The indeterminacy arises because its invariant probability $\frac{\delta}{\delta+\epsilon}$ in the limit becomes a "$\frac{0}{0}$" form indeterminate. This sensitivity to errors likely contributes to the development of automation with different design philosophies by prominent companies.

For instance, consider autopilot systems in aviation, which have become increasingly sophisticated, capable even of landing airplanes under certain conditions. Should these systems be allowed to override human pilots' commands if they deem them unsafe? It's evident that autopilot systems are not infallible. When an autopilot system overrides a correct pilot input, it corresponds to a type-1 error in the model.[2]

Airbus pioneered the "fly-by-wire" technology in commercial aviation, where the ultimate decision-making authority is vested in the flight computer. In these systems, human pilots' control inputs are monitored by the autopilot system, which is designed to override human inputs that exceed safety limits ("flight envelope"). In contrast, Boeing historically maintained a design philosophy where humans retained ultimate control. However, with advances in technology, Boeing has begun

[2]A notorious case is the Maneuvering Characteristics Augmentation System (MCAS) in Boeing 737 Max aircraft, which contributed to two fatal crashes (Lion Air Flight 610 and Ethiopian Airlines Flight 302) due to erroneous angle of attack data, resulting in overrides of pilots' inputs to save the planes.

adopting fly-by-wire technology in newer aircraft like the Boeing 777.[3]

In addition, Corollary 2 also suggests that a machine should not be given final authority if it cannot correct a human mistake ($e_{2,m} = 100\%$ or equivalently $\delta = 0$) and can only execute imperfectly ($\epsilon = e_{1,m} > 0$), despite potentially superior execution capability compared to humans ($e_{1,m} \ll e_{1,h}$). This feature highlights the importance of troubleshooting capabilities for technologies — a distinct feature for generative AI.

Implications for Generative AI

In contrast to traditional machines, generative AI is akin to humans in its ability to produce useful solutions without a correct input, or even without predefined notions of correctness. In the model, this feature can be interpreted as a reduced type-2 error that is bounded away from 100% (or $\delta \nrightarrow 0$). If it can maintain good execution of correct inputs (low type-1 error $\epsilon \to 0$), its invariant distribution $\frac{\delta}{\delta+\epsilon}$ approaches 1. Therefore, such AI is efficient as the ultimate decision maker, offering a final round of error correction without introducing significant new errors. Formally,

Corollary 3 *A generative AI with $\epsilon \to 0$ and $\delta \nrightarrow 0$ should be positioned as the final decision maker. Repeated iterations of AI lead to the probability of correct actions approaching $\frac{\delta}{\delta+\epsilon} \to 1$.*

The evolution from traditional cruise control to self-driving capabilities in automobiles exemplifies this AI transition. Traditional cruise control lacks the ability to steer away from crashes ($\delta = 0$), hence it should never hold the final decision-making authority. In contrast, advanced self-driving systems in latest vehicle models can actively detect and respond to imminent dangers such as rapidly approaching obstacles or other vehicles, often without human intervention ($\delta \nrightarrow 0$). These systems also inherit the executional capabilities of traditional cruise control ($\epsilon \to 0$). Therefore, it is optimal to grant self-driving functions greater authority, including the override of human inputs in certain scenarios.

It's important to note that for generative AI to assume the final decision-making role, it doesn't necessitate AI to surpass humans in troubleshooting (lower type-2 error $e_{2,m} = 1 - \delta < e_{2,h}$). Even though humans may excel in recognizing dangers and avoiding crashes, delegating the final decision to a self-driving car can be optimal because it provides a last line of defense against crashes without

[3]For an interesting comparison between the two design philosophies, see Kornecki and Hall 2004.

introducing new risks. This argument hinges on the assumption that AI maintains high execution quality.

In practice, attempts to solve problems (reduce type-2 error) may inadvertently lead to new errors $\epsilon \nrightarrow 0$, such as AI hallucination. In this case, the invariant probability (10), or equivalently, the ratio $\frac{\delta}{\epsilon}$, provides a measurement for technological advance.

Another crucial feature of AI is its self-learning capabilities based on past performance. Broadly interpreted, this mirrors repeated applications of its own technology in the model. Equation (12) indicates that the posterior probability

$$p_i = \frac{\delta}{\delta + \epsilon} - (1 - \delta - \epsilon)^i \left(\frac{\delta}{\delta + \epsilon} - p_0 \right).$$

converges exponentially towards $\frac{\delta}{\delta+\epsilon} \to 1$ at a rate of $1-\delta-\epsilon$. This feature underscores the explosive growth driven by generative AI.

4 Impact of Technology on Human Effort

4.1 Single Human Layer

In my analysis so far, I've focused on the efficient integration of exogenous technologies. Now, I investigate how technological integration affects human effort incentives to reduce errors. Like machines, humans can make both type-1 and type-2 errors. However, unlike machines, humans can exert costly effort to reduce these errors. The extent of their effort crucially depends on how humans are integrated into the decision-making process. Moreover, human incentives to reduce type-1 and type-2 errors differ significantly.

To introduce humans into the model, consider a strategic player among the N layers. If this player operates in layer $j \in \{1, 2, ..., N\}$, he creates type-i error $h_i(f_{i,j})$ based on his effort input $f_{i,j}$, where $i = 1, 2$. If the correct action is chosen, the player receives a normalized positive utility of 1; otherwise, he receives no utility. The cost of effort is denoted by $c_i(f_{i,j})$ for $i = 1$ or 2. Standard assumptions apply to the error and cost functions:

Assumptions:

- Errors $h_i(f_{i,j})$ are decreasing and convex in $f_{i,j}$ (reflecting decreasing marginal benefit of effort).

- Costs $c_i(f_{i,j})$ are increasing and convex in $f_{i,j}$ (reflecting increasing marginal cost of effort).

The remaining $N-1$ technologies $\mathcal{M}_n$ $(n \neq j)$ are given exogenously and are efficiently arranged by the increasing sequence of invariant probabilities p_n^*, as specified by Proposition 1.

To optimize their effort levels $f_{1,j}$ and $f_{2,j}$ (which affect $\mathcal{M}_j$), the player operating in layer j among the N layers maximizes the expected payoff net of effort costs as follows:

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

It is useful to explicitly derive the first-order conditions with respect to $f_{1,j}$ and $f_{2,j}$.

The first order condition with respect to $f_{1,j}$ (effort on type-1 error) can be decomposed as follows:

$$\begin{aligned}
c_1'(f_{1,j}^*) &= \frac{\partial \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{N} \mathcal{M}_i \mathbf{U}'}{\partial f_{1,j}} \\
&= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} h_1'(f_{1,j}) \prod_{i=j+1}^{N} \mathcal{M}_i \mathbf{U}' \\
&= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \mathbf{U}' \cdot \begin{pmatrix} -1 & 1 \end{pmatrix} h_1'(f_{1,j}) \prod_{i=j+1}^{N} \mathcal{M}_i \mathbf{U}' \\
&= -h_1'(f_{1,j}^*) \cdot \underbrace{\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \mathbf{U}'}_{\text{relevance of type-1 error}} \cdot \underbrace{\prod_{i=j+1}^{N} (e_{2,i} - e_{1,i})}_{\text{consequence of type-1 error}} .
\end{aligned} \tag{16}$$

From equation (16), it is clear that the marginal benefit of effort $f_{1,j}$ operates through two intuitive channels. The first term $\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \mathbf{U}'$ represents the probability of a *correct* action being proposed to the human player in layer j. This is when type-1 error is relevant, and I therefore label this term as the “relevance” of type-1 error in layer j.

The second term $\prod_{i=j+1}^{N} (e_{2,i} - e_{1,i})$ measures the importance of making a correct decision in layer j and is therefore termed the "consequence" of a (type-1) error. Mathematically, it is calculated as the difference between the ultimate correct probabilities p_N conditional on whether a correct or incorrect action is proposed by the human player in layer j.

The first order condition with respect to $f_{2,j}$ (effort on type-2 error) in layer j can be similarly decomposed into the relevance and consequence components:

$$
\begin{aligned}
c_2'(f_{1,j}^*) &= \frac{\partial \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{N} \mathcal{M}_i \mathbf{U}'}{\partial f_{2,j}} \\
&= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} h_2'(f_{2,j}) \prod_{i=j+1}^{N} \mathcal{M}_i \mathbf{U}' \\
&= \begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \end{pmatrix} h_2'(f_{2,j}) \prod_{i=j+1}^{N} \mathcal{M}_i \mathbf{U}' \\
&= -h_2'(f_{2,j}^*) \underbrace{\begin{pmatrix} p_0 & 1-p_0 \end{pmatrix} \prod_{i=1}^{j-1} \mathcal{M}_i \begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\text{relevance of type-2 error}} \underbrace{\prod_{i=j+1}^{N} (e_{2,i} - e_{1,i})}_{\text{consequence of type-2 error}}
\end{aligned}
\tag{17}
$$

Similar to equation (16), the first part in equation (17) represents the relevance of type-2 error in layer j, indicating the probability of an *incorrect* action being proposed by the $j-1$th technology. Comparing equations (16) and (17), note that the consequence terms coincide, reflecting that the effect of human errors in layer j on the final outcome is the same, regardless of which type of error is made.

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

$$
\text{effort incentive}_i \equiv \frac{c_i'(f_{i,j}^*)}{-h_i'(f_{i,j}^*)}, \tag{18}
$$

which, in equilibrium, equals the marginal cost per unit of error reduction. From equations (16) and (17), the effort incentive is the product of the relevance and consequence of each type of error.

With the decomposition and intuition, I introduce the main result of this section, which summarizes how human effort is affected by technological integration.

Proposition 2 ***[Effort Incentive]*** *Suppose a human player is integrated as layer $j \leq N$ in an otherwise optimally arranged sequence of $N-1 \geq 1$ technologies. The player's optimal effort levels $f_{1,j}^*$ and $f_{2,j}^*$ to reduce both types of errors increase with j. Furthermore, the equilibrium effort incentive for reducing type-1 errors increases more quickly than that for type-2 errors. Mathematically,*

$$
\frac{-c_1'(f_{1,j+1}^*)/h_1'(f_{1,j+1}^*)}{-c_1'(f_{1,j}^*)/h_1'(f_{1,j}^*)} > \frac{-c_2'(f_{2,j+1}^*)/h_2'(f_{2,j+1}^*)}{-c_2'(f_{2,j}^*)/h_2'(f_{2,j}^*)} > 1. \tag{19}
$$

Intuitively, the relevance of type-1 errors increases later in the chain because the proposed action is more likely to be correct (Corollary 1). Moreover, the consequence of an error is more severe late in the decision-making process due to fewer subsequent layers that can potentially correct the error. Hence, the effort to correct type-1 error $f^*_{1,j}$ increases with j.

The intuition is more intricate for type-2 errors. Despite the increasing consequence of a type-2 error with j, its relevance decreases along the chain, creating potential ambiguity for type-2 effort incentives. In the appendix, I demonstrate that the effect due to increased consequence (by a factor of $\frac{1}{(e_{2,j}-e_{1,j})}$) dominates the decrease in relevance (by a factor of $\frac{1-p_{j-1}}{1-p_j}$). Hence, the effort to correct type-2 error $f^*_{2,j}$ also increases with j, despite the ambiguity.

The asymmetry in the relevance of type-1 and type-2 errors also leads to the observation that the effort incentive for type-1 errors increases more rapidly than that for type-2 errors.

Implications for AI Technology on Human Effort

Proposition 2 examines the impact of technological integration on human effort. A common concern, frequently highlighted in news reports, is that human pilots or drivers may lose focus or even fall asleep after the autopilot or auto-drive system takes over. Notable incidents include Northwest Airlines Flight 188 in 2009 and Batik Air Flight 6723 in 2024. Such occurrences are even more prevalent with cars equipped with auto-driving features, as multiple cases of Tesla drivers falling asleep at the wheel have been reported. While this negligence is both dangerous and illegal, it is a natural consequence predicted by Proposition 2: Technologies capable of making ultimate decisions reduce the incentive for human effort.

The prediction that individuals will focus more on troubleshooting or innovation (reducing type-2 errors) when technology assumes final execution also has empirical support. A notable example is the US Airways Flight 1549 incident. After a bird strike shortly after takeoff from LaGuardia Airport in New York, the Airbus A320 lost both engines, but nonetheless successfully ditched in the Hudson River. The autopilot system maintained the plane's basic stability (minimizing type-1 execution errors), allowing the pilot to concentrate on troubleshooting the problem and devising a solution (reducing type-2 errors).

Similarly, the recent development of ChatGPT enables software developers to focus more on algorithm design, a more innovative task, while delegating the more tedious coding execution to generative AI.

4.2 Multiple Players and Effort Specialization

The insight from Proposition 2 extends to scenarios involving multiple layers of human decision-making: early layers do not exert effort to reduce errors; middle layers specialize in reducing type-2 errors; and final layers focus on minimizing type-1 errors. To formalize these insights succinctly, I make a slight modification to the effort technology.

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. I assume that each player can either use their status-quo technology or choose to specialize in reducing either type-1 or type-2 errors. By incurring a cost c, players can 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 the type-1 error, and the type-1 error remains positive after the error-reducing effort. Furthermore, to simplify the exposition, I assume that the status-quo technology $\{e_1, e_2\}$ improves the probability of the correct outcome. Using Proposition 1 and (10), this condition can be explicitly written as

$$p_0 < \frac{1 - e_2}{1 - e_2 + e_1}. \tag{20}$$

As before, each player receives 1 unit of utility 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 profile in a Nash equilibrium, where no player has an incentive to deviate given the effort pattern of other players.

Proposition 3 ***[Effort Specialization]*** *All equilibria are characterized by two cutoffs $0 \leq N_1 \leq N_2 \leq N$, such that players in the initial N_1 layers do not make 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.

Using the equilibrium effort pattern and equation (12), I can explicitly calculate the posterior probabilities

$$p_j = \begin{cases} p^*_{(0)} - (e_2 - e_1)^j \left(p^*_{(0)} - p_0\right) & j \leq N_1 \\ p^*_{(2)} - (e_2 - e_1 - \Delta)^{j-N_1} \left(p^*_{(2)} - p_{N_1}\right) & N_1 < j \leq N_2 \\ p^*_{(1)} - (e_2 - e_1 + \Delta)^{j-N_2} \left(p^*_{(1)} - p_{N_2}\right) & j > N_2 \end{cases}, \tag{21}$$

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 errors, and the ones after the reduction of type-1 and type-2 errors, respectively.

The cutoff between the specialization in two types of errors, N_2, is given by

$$N_2 = \max\{j | p_j \leq \frac{1}{2}\}. \tag{22}$$

When the prior probability of being correct exceeds $\frac{1}{2}$, type-1 errors become more relevant than type-2 errors, and players start to specialize in reducing type-1 errors. The other cutoff between effort and no effort, N_1, is determined by

$$N_1 = \max\left\{j | c \leq \min\left\{p_{j-1}\Delta\left(e_2 - e_1 + \Delta\right)^{N-j}, \left(1 - p_{j-1}\right)\Delta\left(e_2 - e_1 - \Delta\right)^{N_2-j}\left(e_2 - e_1 + \Delta\right)^{N-N_2}\right\}\right\}. \tag{23}$$

The first term under the min function in (23) represents the difference in payoffs between no effort and effort to reduce type-1 errors, and the second term is the difference between no effort and type-2 error reduction.

Note that it is possible for some of the three effort regions described in Proposition 3 to vanish. For example, if $N_2 = N$, then no players reduce type-2 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 the N_1th layer specialize in reducing type-2 errors, and no players reduce type-1 errors.

The intuition behind Proposition 3 closely follows from Proposition 2. Early in the decision-making chain, the incentive to exert effort is weaker, so the initial players do not make an effort. As we move to the middle layers, the effort incentive increases; however, type-2 errors are more relevant at this stage because the proposed action is more likely to be incorrect ($p_j < \frac{1}{2}$). Therefore, players in these middle layers focus on reducing type-2 errors. In contrast, in the final layers, the probability of the proposed action being correct increases, making type-1 errors more relevant. Players in these layers specialize in reducing type-1 errors.

Implications for Multi-layered Committee

Revisiting the academic promotion application in Section 2.2.3, Proposition 3 provides insights into the division of labor among different layers of the decision-making process.

The initial layers, such as departmental or school committees, should focus on eliminating type-2 errors. At this stage, a suitable candidate may be incorrectly judged by reference letters, or an unsuitable candidate may be overrated. Since the input is noisy at this stage, the relevance of type-2 errors is high. Conversely, the final layers, such as the provost or president's offices, should focus on eliminating type-1 errors, ensuring efficient execution by promoting promising candidates so they can be retained and removing unsuitable candidates without attracting labor lawsuits. At this stage, the recommendations from the initial layers are likely to be correct, making type-2 errors less relevant. In practice, we often observe that higher-level decision-makers follow the recommendations of the school and department and focus on ensuring that correct procedures are being followed. As such, the promotion (or the denied) cases can be executed efficiently.

Furthermore, the model offers predictions on the length of the decision-making process. Normalize the welfare generated by a correct decision to 1. Suppose each decision-maker introduces a small cost ϵ to the system, such as the communication cost of passing the recommendation to the next layer. The total welfare is then given by

$$p_N - c\,(N - N_1) - \epsilon N.$$

The marginal benefit of a longer decision-making process is its improvement on p_N, which decreases with N. Indeed, one can see from 21 that the marginal benefit decreases exponentially. By comparison, the marginal cost is linear in N, resulting in an optimal interior length for the decision-making process.

5 Multiple Types and Loan Screening Application

In this section, I consider the general model in Section 2.1 with multiple fundamental types $T \geq 2$ and actions $A \geq 2$. The uncertainty in fundamental types is particularly relevant in the loan screening application discussed in Section 2.2.2. Despite being a natural extension of the simplest case, the problem becomes significantly more complex. I begin by examining the case with $A = 2$ actions and multiple fundamental types $T \geq 2$.

Consider a generic technology in (3) for $A = 2$ and an arbitrary $T \geq 2$. Denote by (t) in the superscripts of the probability $\mathbf{P}^{(t)}|_{1\times 2}$, transition matrix $\mathcal{M}^{(t)}|_{2\times 2}$, and payoff vector $\mathbf{U}^{(t)}|_{1\times 2}$ the corresponding values for each fundamental type $t = 1, 2, ..., T$. Using the block diagonal form of $\mathcal{M}$, the payoff (5) can be explicitly rewritten as

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

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,

$$\mathbf{P}_N \mathbf{U}' = \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)} \right) \mathbf{P}_0^{(t)} \prod_{n=1}^{N} \mathcal{M}_n^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mathbf{P}_0^{(t)} \begin{pmatrix} \underline{u}^{(t)} \\ \underline{u}^{(t)} \end{pmatrix}. \tag{24}$$

Intuitively, each fundamental type is associated with a minimum payoff of $\underline{u}^{(t)}$. Hence, the second term in (24) is a constant with respect to different integrations of $\mathcal{M}_n$. Technologies $(\mathcal{M}_n)$ act on the payoff difference between the two actions. Thus, by factoring out $\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)$, the payoff vector can be normalized to $(1, 0)$.

Next, denote by $|\mathbf{P}^{(t)}| \equiv p_1^{(t)} + p_2^{(t)}$ the likelihood of fundamental state t, which does not vary when decision technology is applied. Redefine $\hat{\mathbf{P}}_0^{(t)} = \frac{\mathbf{P}_0^{(t)}}{|\mathbf{P}_0^{(t)}|}$ to be the distribution of initial actions conditional on fundamental type t. The optimal integration problem is equivalent to

$$\mathbf{P}_N \mathbf{U}' = \max_{\sigma(\cdot), I} \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)} \right) |\mathbf{P}_0^{(t)}| \cdot \hat{\mathbf{P}}_0^{(t)} \prod_{i=1}^{I} \mathcal{M}_{\sigma^{-1}(i)}^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{25}$$

To begin, consider $N = 2$ technologies. For each fundamental type 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 directly calculate the difference in payoffs associated with the two orders of applying the technologies:

$$\mathbf{P}_0 \mathcal{M}_1 \mathcal{M}_2 \mathbf{U}' - \mathbf{P}_0 \mathcal{M}_2 \mathcal{M}_1 \mathbf{U}' = \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)} \right) |\mathbf{P}_0^{(t)}| \det \begin{pmatrix} e_{1,1}^{(t)} & e_{1,2}^{(t)} \\ 1 - e_{2,1}^{(t)} & 1 - e_{2,2}^{(t)} \end{pmatrix}, \tag{26}$$

where det denotes the determinant of a matrix. The detailed calculation is relegated to the proof of the next proposition in the Appendix. Importantly, the difference is independent of the initial distribution of actions $\mathbf{P}_0$, but only depends on the fundamental type distribution. This property allows one to rank any two technologies. Formally, we have the following result.

Proposition 4 *Suppose $A = 2$ and $N = 2$. For any $T \geq 1$, a given distribution of fundamental states $|\mathbf{P}_0^{(t)}|$, and the payoff vector $\mathbf{U}$, it is more efficient to apply technology $\mathcal{M}_2$ after $\mathcal{M}_1$ if and only if*

$$\sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)} \right) |\mathbf{P}_0^{(t)}| \det \begin{pmatrix} e_{1,1}^{(t)} & e_{1,2}^{(t)} \\ 1 - e_{2,1}^{(t)} & 1 - e_{2,2}^{(t)} \end{pmatrix} \geq 0. \tag{27}$$

It is useful to note that Proposition 4 nests Lemma 1 and Proposition 1 as a special case with $T = 1$. In this case, condition (27) reduces to

$$\frac{e_{1,1}}{1 - e_{2,1}} \geq \frac{e_{1,2}}{1 - e_{2,2}},$$

which, using condition (10), can be easily verified to be $p_2^* \geq p_1^*$.

When applied to the loan screening application in Section 2.2.2, condition (27) can be explicitly written as

$$r\pi_G \det \begin{pmatrix} e_{1,1}^{(G)} & e_{1,2}^{(G)} \\ 1 - e_{2,1}^{(G)} & 1 - e_{2,2}^{(G)} \end{pmatrix} + L\pi_B \det \begin{pmatrix} e_{1,1}^{(B)} & e_{1,2}^{(B)} \\ 1 - e_{2,1}^{(B)} & 1 - e_{2,2}^{(B)} \end{pmatrix} \geq 0.$$

This highlights two important factors for the lending decision: the credit risk $L\pi_B$ – the risk of lending to a bad type and the associated loss – and the business scope $r\pi_G$ – the risk of failing to lend to a good type and thereby losing interest income.

The comparison between a human lending officer and an automated approval system boils down to their ability to correct errors $(1 - e_2)$ relative to the new errors they introduce (e_1), weighted by the cost of errors associated with each type of borrower.

Lack of Transitivity and Disruption of a New Technology

More interestingly, unlike the case of $T = 1$, the binary relation between two general technologies $(T \geq 2)$ given by Proposition 4 does not possess transitivity. Formally, denote by $\mathcal{M}_2 \succ \mathcal{M}_1$ if condition (27) strictly holds, meaning that it is strictly better to apply $\mathcal{M}_2$ after $\mathcal{M}_1$ when only these two technologies are available. Lack of transitivity means that it is possible for $\mathcal{M}_3 \succ \mathcal{M}_2$ and $\mathcal{M}_2 \succ \mathcal{M}_1$, but $\mathcal{M}_1 \succ \mathcal{M}_3$.

As a counter example for transitivity, consider $\mathbf{U} = (3, 2, 1, 0)$, $\mathbf{P}_0 = (0, 0.5, 0, 0.5)$, and the following three technologies

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

One can verify that the following three relations hold:

$$\begin{aligned} \mathbf{P}_0\mathcal{M}_1\mathcal{M}_2\mathbf{U}' &= 1.415 > 1.41 = \mathbf{P}_0\mathcal{M}_2\mathcal{M}_1\mathbf{U}', \\ \mathbf{P}_0\mathcal{M}_2\mathcal{M}_3\mathbf{U}' &= 1.35 > 1.345 = \mathbf{P}_0\mathcal{M}_3\mathcal{M}_2\mathbf{U}', \\ \mathbf{P}_0\mathcal{M}_3\mathcal{M}_1\mathbf{U}' &= 1.3875 > 1.38 = \mathbf{P}_0\mathcal{M}_3\mathcal{M}_2\mathbf{U}'. \end{aligned} \tag{28}$$

The lack of transitivity also implies that the most efficient integration with more than two technologies becomes intricate. In this numerical example, the most efficient integration with all three technologies is

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

which dominates the other five permutations. However, condition (28) implies that in the absence of $\mathcal{M}_1$, the most efficient way to integrate the remaining two technologies is to apply $\mathcal{M}_2$ first. Therefore, the introduction of a new technology ($\mathcal{M}_1$) can significantly disrupt how existing technologies ($\mathcal{M}_2$ and $\mathcal{M}_3$) are integrated. This feature does not arise when there is no fundamental uncertainty ($T = 1$), because the invariant probability p^* ranks all technologies uniformly.

To summarize, we have the following result.

Proposition 5 *Suppose $T, N \geq 2$ and $\mathcal{M}_n$, $n = 1, 2, ..., N$, is a sequence of optimally integrated technologies, i.e. $\sigma_N^*(n) = n$. Consider the optimal integration $\sigma_{N+1}^*(\cdot)$ with a new technology $\mathcal{M}_{N+1}$. It is possible that σ_{N+1}^* when restricted to $\{1, 2, ..., N\}$ is no longer monotonic.*

Intuitively, the introduction of a new technology alters the relevance and consequence of errors in all layers, thereby significantly affecting the value of existing technologies. This change can shift the optimal order in which technologies should be applied, reflecting the complex interactions between different types of errors and the effectiveness of each technology in addressing them. Consequently, the most efficient integration may vary, leading to a non-monotonic reordering of the technologies even within the set of previously optimal technologies.

6 Technical Extensions

6.1 Higher type-1 error: $e_1 > e_2$

Throughout this work, I have assumed that correcting a mistake is more difficult than maintaining a correct state, i.e., $e_2 > e_1$, which is arguably the more relevant case empirically. In this extension, I consider the possibility when the opposite may be true: $e_1 > e_2$, meaning that correcting a mistake is easier than maintaining a correct action. Intriguingly, the optimal integration of technologies could feature a seemingly bizarre pattern: it may be optimal to utilize a "bad" technology initially to degrade the quality of the proposed action, and then apply a technology with a low type-2 error to correct these mistakes, thereby achieving a superior outcome.

As a simple and extreme example, suppose there are two technologies with the following error patterns:

$$\mathcal{M}_1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } \mathcal{M}_2 = \begin{pmatrix} 1 - e_{1,2} & e_{1,2} \\ 1 & 0 \end{pmatrix}.$$

Essentially, the first technology always creates errors with $e_1 = e_2 = 100\%$. The second can always correct mistakes $e_2 = 0\%$ and its type-1 error is therefore irrelevant ($e_1 > e_2 = 0$). The invariant probability of $\mathcal{M}_1$ is worse than the prior $p_1^* = 0 < p_0$, but it is easy to see that the combination of these two technologies delivers the correct outcome with 100% probability:

$$\begin{pmatrix} p_0 & 1 - p_0 \end{pmatrix} \mathbb{M}_1 \mathbb{M}_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_0 & 1 - p_0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1.$$

This example demonstrates that even when initial actions degrade the quality, subsequent corrections can lead to an optimal outcome, highlighting the complex and sometimes counterintuitive nature of integrating multiple decision technologies.

7 Conclusion

Using linear algebra, I construct a framework to analyze the optimal integration of humans and technologies when both can make two types of errors: changing a correct action (type-1 error) and accepting a wrong action (type-2 error). I first identify the invariant probability of a technology's transition matrix as its quality measure. The optimal integration utilizes technologies that are superior to the prior probability and applies them in ascending order of quality. I then show that human effort to reduce errors increases along the decision-making process, with effort to reduce type-1 errors increasing more rapidly than effort to reduce type-2 errors. Early decision-makers should focus on reducing type-2 errors, while the final decision-makers should focus on reducing type-1 errors. Finally, when multiple fundamental states affect the cost of errors, I construct an explicit condition to optimally integrate two arbitrary technologies. However, this binary relation does not possess transitivity, making a general ranking among more than two technologies infeasible.

This analysis offers insights on when humans or machines should have ultimate authority, how generative AI differs from traditional technologies and why it has such a significant impact, and how integration affects human effort to mitigate errors. Applications include automation design, multi-layered committees, and loan screening.

Finally, I believe that the framework developed in this paper can be utilized to study other questions. For example, consider sequential delegation of a task. Agents have more information about the viability of the task (measuring skill) but are simultaneously biased towards implementing the task (measuring loyalty). Suppose agents have different degrees of skill and loyalty. How does the delegation chain look? How does it affect the types of tasks being delegated? A modified version of the model may shed light on these questions and generate implications for job designs within an firm and corporate governance. I look forward to more research in this area.

References

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

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

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

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

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

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

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

[8] 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.

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

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

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

Appendix

Proof of Lemma 1: Note that both sides of (14) 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 (12) that

$$\begin{pmatrix} p_1^* & 1-p_1^* \end{pmatrix} \mathbb{M}_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in (p_1^*, p_2^*) \,. \tag{29}$$

Apply (12) again, the lower bound of p_1^* in (29) implies that

$$\begin{pmatrix} p_1^* & 1-p_1^* \end{pmatrix} \mathbb{M}_2\mathbb{M}_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} < \begin{pmatrix} p_1^* & 1-p_1^* \end{pmatrix} \mathbb{M}_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p_1^* & 1-p_1^* \end{pmatrix} \mathbb{M}_1\mathbb{M}_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} .$$

Next, consider $p_0 = 0$. Our objective (14) 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 $\mathbb{M}_i$ can be simultaneously diagonalized as follows

$$\begin{pmatrix} 1 & 1-p_i^* \\ 1 & -p_i^* \end{pmatrix} = \begin{pmatrix} p_i^* & 1-p_i^* \\ 1 & -1 \end{pmatrix}^{-1}$$

$$\mathbb{M}_i = \begin{pmatrix} 1 & 1-p_i^* \\ 1 & -p_i^* \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e_{2,i}-e_{1,i} \end{pmatrix} \begin{pmatrix} p_i^* & 1-p_i^* \\ 1 & -1 \end{pmatrix}.$$

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

Proof of Proposition 1: I first establish the following lemma.

Lemma 2 *For any integration σ of N technologies, the final probability p_N is strictly increasing in p_0.*

Proof of Lemma 2: It is clear from (12) 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 efficient integration. Suppose otherwise, the nth technology $\mathbb{M}_n$ 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 (12), we know that the posterior probability in the integration after applying $N_0 - 1$ technologies is higher than p_0. Hence, removing technology $\mathbb{M}_n$ strictly increases the posterior probability after N_0 technologies. Lemma 2 then implies that the final probability is also strictly higher, contradicting with the efficiency of σ^*. Hence, all technologies in the efficient integration must feature invariant probabilities higher than p_0.

In addition, the efficient integration must include all technologies with invariant probability greater than p_0. Otherwise, suppose $\mathbb{M}_n$ 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 2.

Next, I show that the efficient 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 efficient 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 more efficient to remove the N_0+1th technology.

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

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

Proof of Proposition 2: 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 (16) and (17) can be rewritten as

$$-p_{j-1}h_1'(f_{1,j}^*)\Pi_{i=j+1}^N(e_{2,i}-e_{1,i}) = c'(f_{1,j}^*), \tag{30}$$

and

$$-(1-p_{j-1})\,h_2'(f_{2,j}^*)\Pi_{i=j+1}^N(e_{2,i}-e_{1,i}) = c'(f_{2,j}^*). \tag{31}$$

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

$$-\left[p_{j-1}(1-e_{1,j})+(1-p_{j-1})(1-e_{2,j})\right]h_1'(f_{1,j+1}^*)\Pi_{i=j+2}^N(e_{2,i}-e_{1,i}) = c'(f_{1,j+1}^*), \tag{32}$$

and

$$-\left[p_{j-1}e_{1,j}+(1-p_{j-1})\,e_{2,j}\right]h_2'(f_{2,j+1}^*)\Pi_{i=j+2}^N(e_{2,i}-e_{1,i}) = c'(f_{2,j+1}^*). \tag{33}$$

Corollary 1 implies that the posterior probabilities p_i are increasing in i. Comparing (30) and (32), I have

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

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 (31) and (33), 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 (19), 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 efficiently 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 3: 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:

$$
\begin{aligned}
c \geq & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \left[\begin{pmatrix} 1-e_1+\Delta & e_1-\Delta \\ 1-e_2 & e_2 \end{pmatrix} - \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2 & e_2 \end{pmatrix} \right] \textstyle\prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} \Delta & -\Delta \\ 0 & 0 \end{pmatrix} \textstyle\prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \textstyle\prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta p_{j-1} \textstyle\prod_{n=j+1}^{N} \left(e_{2,n} - e_{1,n} \right).
\end{aligned}
\tag{34}
$$

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:

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

The strict inequality holds because $p_{j'-1} < p_{j-1}$ and $\prod_{n=j'+1}^{j} \left(e_{2,n} - e_{1,n} \right) < 1$. The former condition arises from the fact that no one in the initial $j'-1$ layers makes effort, and condition (20) 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 (34) 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 compatibility condition for layer-j player implies that

$$\begin{aligned}
c \geq & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \left[\begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2+\Delta & e_2-\Delta \end{pmatrix} - \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2 & e_2 \end{pmatrix} \right] \textstyle\prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ \Delta & -\Delta \end{pmatrix} \textstyle\prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \textstyle\prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
= & \Delta \left(1-p_{j-1}\right) \textstyle\prod_{n=j+1}^{N} \left(e_{2,n}-e_{1,n}\right).
\end{aligned} \tag{35}$$

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 (35) 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

$$
\begin{aligned}
0 \geq\ & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \left[\begin{pmatrix} 1-e_1+\Delta & e_1-\Delta \\ 1-e_2 & e_2 \end{pmatrix} - \begin{pmatrix} 1-e_1 & e_1 \\ 1-e_2+\Delta & e_2-\Delta \end{pmatrix} \right] \prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
=\ & \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} \Delta & -\Delta \\ -\Delta & \Delta \end{pmatrix} \prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
=\ & \Delta \begin{pmatrix} p_{j-1} & 1-p_{j-1} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \prod_{n=j+1}^{N} \mathbb{M}_n \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
=\ & \Delta \left(2p_{j-1}-1\right) \prod_{n=j+1}^{N} \left(e_{2,n}-e_{1,n}\right),
\end{aligned}
\tag{36}
$$

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 (36) 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

$$p^*_{(0)} < p^*_{(2)} \leq p^*_{(1)}. \tag{37}$$

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 (37) 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 4: 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 (12),

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

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 (25) becomes

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

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

$$\sum_{t=1}^{T}\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)|\mathbf{P}_0^{(t)}|\left[\frac{1}{p_1^{(t)*}} - \frac{1}{p_2^{(t)*}}\right]\left(1 - e_{2,2}^{(t)}\right)\left(1 - e_{2,1}^{(t)}\right) \geq 0.$$

Together with

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

The above condition becomes

$$\sum_{t=1}^{T}\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)|\mathbf{P}_0^{(t)}|\left[\frac{e_{1,1}^{(t)}}{1 - e_{2,1}^{(t)}} - \frac{e_{1,2}^{(t)}}{1 - e_{2,2}^{(t)}}\right]\left(1 - e_{2,2}^{(t)}\right)\left(1 - e_{2,1}^{(t)}\right) \geq 0.$$

After manipulation, it becomes

$$\sum_{t=1}^{T}\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)|\mathbf{P}_0^{(t)}|\left[e_{1,1}^{(t)}\left(1 - e_{2,2}^{(t)}\right) - e_{1,2}^{(t)}\left(1 - e_{2,1}^{(t)}\right)\right] \geq 0,$$

establishing conditions (26) and (27). ■