Optimal Integration: Human, Machine, and Generative AI

Hongda Zhong, University of Texas at Dallas and CEPR*

August 10, 2024

Abstract

I study the optimal integration of human and technologies in multi-layered decision-making processes. Each layer can correct existing errors but may simultaneously introduce new errors. The ratio of these two probabilities is a one-dimensional quality metric, offering an optimal integration rule: deploying higher quality technologies in later stages. Technologies influence human effort asymmetrically: early layers exert lower effort and emphasize error correction, whereas later layers exert higher effort and focus on avoiding new errors. Extending the model to multiple actions and states, no single metric can universally determine the optimal integration rule. Therefore, new technologies can significantly disrupt existing integration.

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

*For valuable discussions and comments, I thank Alain Bensoussan, Jonathan Berk, Mike Burkart, Philip Bond, Will Cong, Ron Kaniel, Ivo Welch, Lucy White, Xingtan Zhang, and the seminar participants at UT-Dallas and CKGSB. All errors are my own.

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, autonomous vehicles, and FinTech lending. While these technologies can often help reduce human errors, they may also introduce new errors into the system. Several questions naturally arise: How should humans and technologies be integrated optimally? 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?

Motivated by these questions, I propose a framework based on linear algebra to study the optimal multi-layered decision process. Each 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, generating payoffs that can potentially depend on some unobservable states.

The model allows for multiple actions and payoff relevant states. For simple illustration, I consider the case of binary actions in the context of automation, such as aviation autopilot or autonomous vehicles. With a great level of abstraction, actions are broadly classified into binary categories: safe or dangerous. A decision maker (human or automative technology) 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 optimal integration maximizes the probability of adopting a safe action, which informs the ideal allocation of authority among human 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 input action distribution by this transition matrix yields the distribution of recommended actions for the subsequent layer. The optimal integration is a particular sequence to multiply these matrices to maximize the probability of ultimately taking the correct action. To my knowledge, there is not a readily available mathematical result that determines the optimal sequence to multiply matrices, making it one the contributions of this paper.

The optimal integration follows a simple rule of thumb: Deploying higher "quality"

technologies in later stages. Intuitively, superior technologies should act as "gatekeepers" in decision-making to prevent subsequent inferior technologies from erroneously altering a safe action, thereby improving outcomes.

However, defining a "quality" metric in multi-dimensional actions (e.g. two types of errors e_1 and e_2) and states is not straightforward. One theoretical contribution of this work is establishing quality metrics for several economically relevant applications and demonstrating the absence of a universal metric in general.

In the simplest case of binary actions, 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 ranking decision technologies uniformly. This metric also reflects the ratio of error correction capability ($1-e_2$) to new errors introduced (e_1). Technologies shift the probability of a safe action towards their invariant probability. Optimal integration progressively increases the likelihood of achieving a safe action throughout the decision-making process.

I then explore how human efforts to mitigate errors are influenced by their placement within the integration. Interestingly, incentives for effort across layers show distinct patterns in reducing type-1 and type-2 errors: While both types of effort increase in later layers, the effort for reducing type-1 error increases more rapidly. In other words, initial agents exert lower effort and focus on reducing type-2 errors (or equivalently, correcting existing errors), and final agents exert higher effort and focus on reducing type-1 errors (or equivalently, avoiding new errors).

Intuitively, two factors impact effort incentives: the relevance of a particular type of error faced by humans and the consequence of human errors on the correctness of the 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, because 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.

In the full model, I consider multiple states and feasible actions to capture various

error types with differing severities. For instance, in loan screening, banks have two actions, lending or rejecting applicants, with payoffs depending on the state of the borrower (good or bad). Lending to bad borrowers can lead to significant costs from potential default losses, while rejecting creditworthy applicants may result in mild losses from foregone interest. Generic technologies can alter proposed actions arbitrarily depending on the state, characterized mathematically by a probability transition matrix in the state-action space (e.g., a 4×4 matrix for the loan screening application).

The general analysis is intricate and yields novel insights. With binary actions and multiple states, for any pair of technologies, I can establish an explicit condition to determine their optimal integration, independent of initial action distribution. This condition incorporates error costs and the relative quality of the two technologies for each state. Surprisingly, this binary relation lacks transitivity when more than two technologies are available. Specifically, 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 the optimal order when only each respective pair is available.

The lack of transitivity implies that no one-dimensional "quality metric" universally exists for generic technologies. It also delivers an interesting insight: The introduction of a new technology may significantly alter existing integration. In the previous example, if $\mathcal{M}_3\mathcal{M}_1\mathcal{M}_2$ is the optimal integration of all three technologies, then the introduction of $\mathcal{M}_1$ changes the relative position of $\mathcal{M}_2$ and $\mathcal{M}_3$.[1] A new technology changes the relative importance of different errors (or equivalently states), which in turn affects the relative quality between existing technologies.

The analysis of multiple actions also confirms the general absence of a one-dimensional quality metric. Moreover, optimal integration may vary with initial action distributions, complicating matters further. I look forward to future research offering insights into these complexities.

Interpretation of Results in Various Applications

[Automation] In the application of automation (or more broadly, binary actions), type-1 error reflects execution quality (the ability to maintain safe actions) whereas

[1]Recall that the optimal integration when only $\mathcal{M}_2$ and $\mathcal{M}_3$ are available is $\mathcal{M}_2\mathcal{M}_3$. It should also be clear that, with circular order, any three-technology integration will result in a reversal of positions for two of the technologies.

type-2 error measures troubleshooting capabilities (the ability to correct dangerous actions). Traditional technologies 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 traditional technologies is therefore highly sensitive to the specific combination of the two errors, explaining varied approaches towards automation among companies.[2]

[Generative AI] Unlike traditional automative technologies, generative AI, such as ChatGPT or autonomous vehicles, occasionally demonstrate troubleshooting capabilities that reduce type-2 errors from near 100%. For instance, autonomous driving technologies can sometimes avert collisions without driver input. Consequently, the invariant probability of generative AI rapidly approaches 100%. The optimal integration rule suggests AI should serve as the ultimate decision-maker for execution.

[Human-AI Collaboration] For AI to take ultimate execution, it does not require AI to outperform humans in troubleshooting, nor does it exclude human involvement. Instead, humans focus (more) on troubleshooting initially, while AI handles final execution. AI's strength lies in enhancing troubleshooting without significantly increasing new errors, maintaining high execution quality (low type-1 error). Therefore, if AI introduces a non-negligible amount of new errors, such as AI hallucinations, the applicability of such a technology is significantly compromised.

[Effort Incentive] The model's prediction on human effort 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.[3] 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.[4]

[Multi-layer Approval Process] A pertinent example of multi-layered decision-making is the academic promotion (tenure) process, which frequently involves mul-

[2]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).

[3]Notable incidents include Northwest Airlines Flight 188 in 2009 and Batik Air Flight 6723 in 2024. See Section 4.1 for details.

[4]A notable example in aviation is the US Airways Flight 1549 incident. Another example is the utilization of generative AI tools by software developers. See Section 4.1 for details.

tiple rounds of evaluation, including assessments from letter writers, departmental committees, schools, provosts, 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 of the earlier rounds. Conversely, higher-level decision-makers such as provosts and presidents rarely overturn school recommendations. Instead, they often focus on efficient execution of the decisions: ensuring procedural adherence and minimum administrative burden to promote successful candidates; and smooth dismissal of unsuccessful candidates without generating adverse consequences for both the university and the candidates.

Similar hierarchical structures are observed in other multi-layer approval processes such as budget reviews, student admissions, and accounting audits. Lower-tier agents prioritize error reduction, while higher-tier decision-makers, with ultimate authority, emphasize procedural fidelity and executional correctness.

Literature Review

This paper contributes to several strands of literature in decision science and economics. 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 linear algebra, they address distinct problems. AHP uses eigenvectors to prioritize attributes based on pairwise preferences, while my analysis attempts to determine the optimal sequence to multiply 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 multi-layer decision-making process with T unobservable fundamental states and A available actions. Let $\mathbb{T}$ and $\mathbb{A}$ denote the sets of states and actions, respectively. The probability distribution of fundamental states is represented by $\mathbf{q} = (q_1, q_2, ..., q_T)$. Action a taken in state t creates a payoff (or utility) to the designer U_{at}, and these payoffs are collectively captured by a real vector

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

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 state t. Together with the fundamental distribution $\mathbf{q}$, the decision induces a probability distribution on the AT dimensional state-action space

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

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

There are $N \geq 2$ decision makers (human or technology) that can be integrated sequentially. Each decision maker 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. The output may be random and follows a distribution on $\mathbb{A}$ depending on the fundamental state. 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 decision maker is mathematically characterized by an $AT \times AT$ dimensional probability transition matrix $\mathcal{M}$. The posterior distribution of the actions proposed by the decision maker is given by $\mathbf{P}_1 = \mathbf{P}_0\mathcal{M}$. It captures arbitrary ways that a decision maker may propose actions based on the underlying state (e.g. from learning about the fundamental state) and the input action from the previous layer. It also allows the possibility that a decision maker can change the fundamental state. I provide specific forms of $\mathcal{M}$ for several different applications in Section 2.2.

A technology's matrix $\mathcal{M}$ is exogenously given – a benchmark case analyzed in Section 3. In contrast, a human decision maker's $\mathcal{M}(\mathbf{f})$ may depend on the his multi-dimensional effort input $\mathbf{f}$ – a case formalized and analyzed in Section 4.

When N decision makers (human or technology) $\{\mathcal{M}_n\}$ are applied sequentially, the posterior distribution of actions is given by

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

and the expected payoff to the designer is therefore

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

An integration of (a subset of) the N decision makers is represented 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 $i = 1, 2, ..., I$. In words, only I out of N technologies are included in the integration, and $\sigma(n)$ denotes the layer index of the original technology n within the integration. Layer I is the final decision maker. A technology is excluded if $\sigma(n) = 0$.

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

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

It is worth noting despite the simple setup, the problem is mathematically difficult. To my knowledge, there is not a general mathematical result for this problem.

2.2 Applications of the Model

2.2.1 Automation – Autopilot and Self-Driving

An example of the simplest case with $T = 1$ and $A = 2$ is automation in the transportation industry (autopilot or autonomous vehicle). Using the example of aviation, with a great level of abstraction, there are 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 $\mathbf{U} = (1, 0)$. The actions' payoffs are not state dependent, or in other words, there is only $T = 1$ state. (For comparison, see loan screening application in Section 2.2.3.)

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 safe action taken by the pilot (or fail to correct a dangerous action). 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. Technologies can failure, such as sensor failures, and in addition, automation might also make human pilots complacent. The model effectively captures these complexities.

A decision maker (human pilots or autopilot) can be generically represented by the following 2×2 matrix

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

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

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

Throughout the discussion, I assume type-2 error is higher than type-1 error ($e_2 > e_1$), that is correcting a mistake is more difficult than maintaining a correct action.

Indeed, 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) 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, and how to divide the final authority between human pilots and autopilots.

2.2.2 Multi-Layer Approval – Academic Promotion

The model can be applied to study multi-layered approval processes, including academic promotion review, budget review, student admission, accounting and auditing in companies.

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 $\mathcal{M}_n(\mathbf{f}_n)$ (details in Section 4) offers insights into the specific aspects that each committee should prioritize. Moreover, the model can also shed light on how automative technologies affect human incentive.

2.2.3 Loan Screening

FinTech (automated loan screening) is a particularly pertinent example when multiple payoff relevant states become crucial. Consider $T = 2$ types (states) 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 $\mathbf{U} = (r, 0, 0, -L)$, based on the four $(A \times T = 4)$ outcomes: accepting good, rejecting good, rejecting bad, accepting bad borrowers $\{GA, GR, BR, BA\}$.

A loan screening technology (human loan officer or automated screening system)

$\mathcal{M}$ is a probability transition matrix on these four outcomes:

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

Since a technology can change lending decisions, but does not alter the borrower type between G and B, the matrix $\mathcal{M}_n$ only contains $T = 2$ diagonal blocks, each of size 2×2 ($A \times A$) – there is no probability transition between good and bad type borrowers. Within each state (borrower type), technologies are determined by their respective type-1 and type-2 errors: $e_{1,n}^{(\theta)}$ 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.

More generally, decision technologies that alter the proposed action, but not the fundamental state take the block diagonal form

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

The state-dependent transition matrices on actions $\mathcal{M}^{(t)}|_{A\times A}$ allow for arbitrary learning of the underlying fundamental state in the decision-making process.

3 Optimal Integration – Simplest Case

I start by analyzing the simplest case with binary actions ($A = 2$) and a single state ($T = 1$), motivated by the application of automative technologies in Section 2.2.1. Throughout this section, I use 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 5), because one can always relabel "correct" action in each state as the one generating higher payoff. The analysis delivers several key insights that serve as the foundation for more general cases and enriched models.

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 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,

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

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

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,

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

Condition (13) 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^*. With this intuition, we are prepared to characterize optimal integration strategies.

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,*

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

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

Intuitively, to achieve a superior outcome, the designer should position the higher "quality" technology later in the sequence as the "gatekeeper." Conversely, placing a good technology early is suboptimal since subsequent technologies introduce more errors and compromise correct actions made by the initial technology. However, it is not clear how one can define "quality" due to multi-dimensional error profiles: one technology may create higher type-1 errors while the other generates higher type-2 errors. Proposition 1 is a powerful result by reducing the dimensionality of the problem and shows that the invariant probability is the key quantity to rank different technologies.

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

Furthermore, thanks to the 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 can be generalized 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 associated with 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 will be particularly relevant when considering human effort in decision-making processes.

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

Before proceeding to the applications of Proposition 2, it's important to note it is not generally possible to extend binary ranking between technologies to arbitrary number of them (for such an example, see the case of $T \geq 2$ in Section 5).

Implications for Automation

Revisiting the application discussed in Section 2.2.1, the benchmark result provides insights into several questions regarding technology and generative AI. Should human or technology have the ultimate authority? What sets generative AI apart from traditional technologies? Why does it have such a profound impact?

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 (e.g. sensor failure which is rare), 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 recovery from the situation when autopilot 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 2 offers implications on 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 optimal integration, i.e.* $\frac{1-e_{2,h}}{1-e_{2,h}+e_{1,h}} > p_0$, *a machine should have final authority if and only if*

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

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.

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 adopting fly-by-wire technology in newer aircraft like the Boeing 777.[5]

In addition, Corollary 2 also suggests that a machine should not be given final authority if it can only execute imperfectly ($\epsilon = e_{1,m} > 0$), but cannot correct a mistake ($e_{2,m} = 100\%$ or equivalently $\delta = 0$). This is true even when the machine's execution capability is better than humans ($e_{1,m} \ll e_{1,h}$). This observation highlights

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

the importance of troubleshooting capabilities for technologies — a distinct feature for generative AI.

Implications for Generative AI

In contrast to traditional machines, generative AI may be able to produce useful solutions without necessarily a correct input, or even without predefined notions of correctness.[6] 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 probability $\frac{\delta}{\delta+\epsilon}$ approaches 1. Therefore, it is optimal to have such an AI as the ultimate decision maker, offering a final round of error correction without introducing significant new errors.

Another important 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 $1-\delta-\epsilon$. The speed of convergence (learning) may be slow especially when δ is small, but the rate is exponential, which potentially explains the explosive improvement of generative AI technologies. Formally,

Corollary 3 *A generative AI with $\epsilon \to 0$ and $\delta \nrightarrow 0$ should be positioned as the final decision maker. Furthermore, 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 autonomous driving capabilities 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

[6]For example, ChatGPT can write a poem without much detailed instructions from human. Similarly, autonomous driving can sometimes steer away from danger without driver inputs.

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 does not necessitate AI to surpass humans in troubleshooting (i.e., AI having a lower type-2 error $e_{2,m} = 1 - \delta < e_{2,h}$). Even though humans may perform better in recognizing dangers and avoiding crashes, delegating the final decision to AI powered vehicle can be optimal because it provides a last line of defense against crashes without introducing new risks.

The argument for AI having ultimate authority hinges on the assumption that AI does not introduce new errors. 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, and high type-1 errors ϵ significantly limits the usefulness of AI.

4 Impact of Technology on Human Effort

4.1 Single Human Decision Maker

The analysis so far focuses on the optimal integration of exogenous technologies. Now, I investigate how technological integration affects human effort incentives to reduce errors. Unlike machines with exogenous error probabilities, humans can exert costly effort to reduce these errors. An intriguing finding from this section is that human effort incentive depends on their location in the integration and behaves asymmetrically for type-1 and type-2 errors.

To introduce human effort, suppose that there is a strategic decision maker among the N layers, henceforth, the "player". 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 $i = 1, 2$. I still consider the case of binary actions as in Section 3, because this setting gives a clear interpretation of "error."[7] I use the notation h_i to highlight the endogenous human errors and distinguish from the exogenous errors of technologies e_i.

[7]Recall that one can always label the action that generates a higher payoff as the correct action. When multiple actions ($A > 2$) are available, the mathematical expressions are similar but the interpretation is less clear.

If the correct action is chosen, the player receives some positive utility; otherwise, he receives no utility. Given the binary nature, the level of positive utility is inconsequential, and I normalize it to 1. 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:

- 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 optimally arranged in the ascending order of their invariant probabilities p_n^*, as specified by Proposition 2. The case with N human players is analyzed in Section 4.2.

To optimize effort levels $f_{1,j}$ and $f_{2,j}$ (which affect $\mathcal{M}_j$), the player operating in layer j 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:

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

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

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 an error in layer j affects the correctness of the final action, and is therefore termed the "consequence" of a (type-1) error. Mathematically, it is the change in 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 of an error in layer j is the same, regardless of its type.

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.[8] From equations (16) and (17), the effort incentive can be decomposed into the relevance and consequence of each type of error.

[8]To see that the expression captures effort incentive, the classic linear-quadratic benefit-cost functions $c(f)=\frac{1}{2}f^2$ and $h(f)=1-f$ (when restricted to the natural probability domain), gives the exact equilibrium effort choice $\frac{c'(f^*)}{-h'(f^*)}=f^*$. More generally, under the natural convexity and monotonicity assumptions, expression (18) is increasing in f^*.

I can now introduce the main result of this section, summarizing how human effort is affected by technological integration.

Proposition 3 ***[Effort Incentive]*** *Suppose a human player operates in layer j of an N-layer decision making process. The other $N-1$ technologies are optimally integrated according to Proposition 2. 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 that for $f^*_{2,j}$. Mathematically,*

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

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 increasing relevance of type-1 errors and decreasing relevance of type-2 errors also imply a more rapid increase in effort incentive towards type-1 errors.

Implications for AI Technology on Human Effort

A common concern of relying on automation, frequently highlighted in news reports, is that human controller (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 autonomous driving, 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 3: Technologies capable of making ultimate decisions reduce the incentive for human effort.

A silver lining of the effort pattern is that human players can now focus more on troubleshooting or innovation (reducing type-2 errors) when technology assumes final execution. A notable example in aviation 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 pilots to concentrate on troubleshooting the problem and devising a solution (reducing type-2 errors) with the extremely limited amount of time (three minutes) they had.

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 3 extends to scenarios involving multiple layers of human decision makers: 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 simplification to the effort technology, making it a discrete choice rather than continuous.

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 type-1 error, and the type-1 error remains positive after its reduction. 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 2 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 pattern in a Nash equilibrium, where no player has an incentive to deviate given the equilibrium effort profile of other players.

Proposition 4 ***[Effort Specialization]*** *All Nash equilibria are characterized by two cutoffs* $0 \leq N_1 \leq N_2 \leq N$*, giving at most three effort regions. 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. In the generic case when all three effort regions exist,[9] I can explicitly calculate the posterior probabilities from equation (12) as follows:

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

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 or type-2 errors, respectively.

In the early stages of the decision-making process, the incentive to exert effort is low, as mistakes here have low consequences. Hence, initial players do not make effort until layer N_1, where

$$N_1 = \max\left\{j | c \leq (1 - p_{j-1})\, \Delta\, (e_2 - e_1 - \Delta)^{N_2 - j}\, (e_2 - e_1 + \Delta)^{N - N_2}\right\}. \tag{22}$$

The right-hand side of the inequality in (22) represents the difference in payoffs between no 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 ($p_j \leq \frac{1}{2}$). Hence, type-2 errors are more relevant.

[9] This case occurs 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.

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

$$N_2 = \max\{j|p_j \leq \frac{1}{2}\}, \tag{23}$$

players start to focus on the more relevant type-1 errors. The overall intuition closely follows and generalizes Proposition 3 to the case of multiple players.

Note that it is possible for some of the three effort regions described in Proposition 4 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 layer N_1 specialize in reducing type-2 errors, and no players reduce type-1 errors.

Implications for Multi-layered Committee

Revisiting the application of academic promotion in Section 2.2.2, Proposition 4 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 error correction (reducing type-2 errors). At this stage, the input is relatively noisy: a suitable candidate may be incorrectly evaluated by reference letters, or an unsuitable candidate may be overrated. Hence, the relevance of type-2 errors is high. Conversely, the final decision makers, such as the provost or president's offices, should focus on execution of the process (reducing type-1 errors), that is to ensure proper procedures are followed. Promising candidates are promoted without too much frictions so they can be retained, and unsuitable candidates are dismissed without attracting labor lawsuits. At this stage, the recommendations proposed by the initial layers are likely to be correct, making the execution aspect more important (reducing type-1 errors).

Similar division of focus is observed in many other multi-layer approval process, for example budget review, student admission, accounting and auditing. Initial layers focus on error detection and correction, whereas final layers focus on the design and execution of the process.

Furthermore, a slightly modified model offers predictions on the optimal length of the decision-making process N. Again, normalize the total welfare generated by a correct action 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

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

From (21), one can see that p_N converges to $p^*_{(1)}$, and therefore the marginal benefit of having a larger N diminishes exponentially. By comparison, the marginal cost is linear in N, resulting in an optimal interior length for the decision-making process.

5 Multiple States and Loan Screening Application

By focusing on the case with $A = 2$ actions and $T = 1$ payoff-relevant state in the baseline model, I effectively assume all errors are equally costly. While the simplification makes the model more intuitive and offers insights more clearly, this is clearly an unsatisfactory feature in practice. For example, in the application of loan screening in Section 2.2.3, accepting a bad borrower is probably significantly more costly than rejecting a good borrower.

In this section, I explicitly consider the possibility that different errors have different consequences, mathematically captured by having 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. When $A \geq 3$, a case analyzed as an extension in Section 6.1, the notion for "correct" or "wrong" becomes less clear.

As will be clear shortly, the problem becomes significantly more complex and new insights are delivered. I first derive an integration rule for $N = 2$ technologies that is akin to Proposition 1. Next, I show that a result similar to Proposition 2 does not exist, that is there does not exist an one-dimensional quality metric that determines an optimal integration rule for $N \geq 3$ technologies.

Consider a generic technology in (8) for $A = 2$ actions and an arbitrary states $T \geq 2$. 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 objects for each fundamental state $t = 1, 2, ..., T$. Using the block diagonal form of $\mathcal{M}$, the payoff (4) can be explicitly rewritten as

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

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 state 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 $\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)$. Factoring it out, the payoff vector can be normalized to $(1, 0)$.

Next, denote by $|\mathbf{P}_0^{(t)}|$ (the sum of its two entries) the probability of fundamental state t, which does not vary along the decision-making process. (Recall that decision technologies only affect proposed actions, but not the fundamental state.) 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 state t. The optimal integration problem is equivalent to

$$\max_{\sigma(\cdot), I} \sum_{t=1}^{T} \left(\overline{u}^{(t)} - \underline{u}^{(t)}\right) |\mathbf{P}_0^{(t)}| \cdot \hat{\mathbf{P}}_0^{(t)} \prod_{i=1}^{I} \mathcal{M}_{\sigma^{-1}(i)}^{(t)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{25}$$

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 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 result in the Appendix. An important feature of (26) is that the difference is independent of the initial distribution of actions $\hat{\mathbf{P}}_0^{(t)}$, but only depends on the fundamental state distribution $|\mathbf{P}_0^{(t)}|$. This property allows one to rank

any two technologies independent of the initial action distribution $\hat{\mathbf{P}}_0^{(t)}$. Formally, we have the following result.

Proposition 5 *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 optimal to apply technology $\mathcal{M}_2$ after $\mathcal{M}_1$ if and only if*

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

It is useful to note that Proposition 5 nests 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 $p_2^* \geq p_1^*$.

Condition (27) 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 ($e_{1,n}^{(t)}$). Therefore, the overall comparison further puts weight on the cost of making a mistake in that state $\left(\overline{u}^{(t)} - \underline{u}^{(t)}\right)$ and the likelihood of each state $|\mathbf{P}_0^{(t)}|$.

When applied to the loan screening application in Section 2.2.3, 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 condition highlights three important considerations when comparing different lending technologies (e.g. loan officers v.s. FinTech algorithms): borrower composition (π_G and π_B), the cost of error for each type (r or L), and the relative performance of the technologies for each borrower type measured by the determinants.

Lack of Transitivity and Disruption of a New Technology

Intriguingly, unlike the case of $T = 1$, the binary relation given by Proposition 5 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. Then, it is possible for $\mathcal{M}_3 \succ \mathcal{M}_2$ and $\mathcal{M}_2 \succ \mathcal{M}_1$, but $\mathcal{M}_1 \succ \mathcal{M}_3$. The lack of transitivity also implies that there is no one-dimensional metric that can rank all technologies, making it impossible to establish a general integration rule like Proposition 2.

For such a counter example, 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:

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

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 the other five permutations. However, condition (28) 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, 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^* is a one-dimensional quality metric that ranks all technologies universally.

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 vary, leading to a non-monotonic reordering of existing technologies.

To summarize, I have the following result.

Proposition 6 *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 not monotonic.*

6 Technical Extensions

6.1 More than Two Actions: $A > 2$

The problem becomes more complex when three or more actions become available ($A \geq 3$). Specifically, the optimal integration may not only depend on the probability of the fundamental states $|\mathbf{P}_0^{(t)}|$, but also the initial distribution on actions $\hat{\mathbf{P}}_0^{(t)}$. Consequently, for different initial actions, the optimal integration sequence may be different. I have yet to find a succinct expression of the payoff difference similar to equation (26).

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

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

Compare two distributions of initial actions $\mathbf{P}_0 = (0, 1, 0)$ and $\hat{\mathbf{P}}_0 = (0, 0, 1)$. One can verify that

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

and

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

Hence, the optimal integration sequence varies with the initial distribution of actions even without fundamental uncertainty.

Intuitively, with $A \geq 3$ actions, the notion of correct or wrong action is no longer as clear. An action a_2 may generate a payoff superior to a_3, but may be dominated by a_1. A technology may be good at improving a_2 to a_1, but bad at handling a_3. Hence, which technology should be applied first can depend on the initial action distribution.

6.2 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 opposite case: $e_1 > e_2$.

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:

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

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: rejecting 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 multi-layered delegation of a project choice to agents (an extension of Aghion and Tirole 1997). Agents have more information about the quality of projects (measuring skill) but may have preferences different from the principal (measuring loyalty). Suppose agents have different degrees of skill and loyalty. Who should be delegated the choice first and who should have ultimate authority to decide? 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] Aghion, Philippe, and Jean Tirole. "Formal and real authority in organizations." Journal of political economy 105.1 (1997): 1-29.

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

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

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

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

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

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

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

[9] 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.

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

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

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

Appendix

Proof of Proposition 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

$$\left(\begin{array}{cc} p_1^* & 1-p_1^* \end{array} \right) \mathcal{M}_2 \left(\begin{array}{c} 1 \\ 0 \end{array} \right) \in (p_1^*, p_2^*) . \tag{29}$$

Apply (12) again, the lower bound of p_1^* in (29) 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 (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 $\mathcal{M}_i$ 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 (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 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 (12), 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: 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

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

and

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

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 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 4: 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:

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

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:

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

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 (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

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

From the definition of p_{j-1}

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

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

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

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

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

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

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

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 5: 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). ■