

A Study of the Prince–Minister Model for AI-Assisted Teaching

Wenjie Chen¹, Xun Zheng¹

¹Shool of Urban Railway Transportation, Shanghai University of Engineering Science, Shanghai, China

University-Level Undergraduate Teaching Construction Project: A Teaching Construction Project on Curriculum Integration Pathways and Practices of the Human-Engaged Computing (HEC) Philosophy "Design of Flow Guidance and AI Integration in Large-Class Teaching Based on HEC Theory" (y202610002)

DOI: <https://doi.org/10.5281/zenodo.23399539>

Published Date: 15-October-2026

Abstract: Artificial intelligence increasingly serves as an advisor in teaching and learning, yet it may be honest or pandering. Users cannot directly observe which type they face. This paper develops the Prince–Minister model to analyze how users should respond to advice from a potentially pandering AI. The user, treated as a prince, faces a binary decision. The AI, treated as a minister, either truthfully reports a private signal or always recommends intervention. The user updates beliefs, chooses an action to minimize expected loss, and may query an independent signal at a cost. The model derives the optimal action rule, query strategy, and equilibrium. It shows that positive AI messages are less trustworthy than negative ones, that querying is most valuable when the decision is uncertain, and that verification is worthwhile only when its benefit exceeds its cost. These results imply that students and teachers should be cautious toward affirmative AI feedback, take critical feedback seriously, verify selectively, and weigh verification costs against benefits. The model offers a decision-theoretic foundation for critical AI literacy, treating AI as an advisor to be evaluated rather than an authority to be obeyed.

Keywords: Prince–Minister model; AI-assisted teaching; pandering AI; advice-taking.

I. INTRODUCTION

Artificial intelligence is increasingly embedded in teaching and learning. Students use AI to check answers, generate explanations, and decide what to study next. Teachers use AI to design materials, assess student work, and make instructional decisions. In many of these settings, AI no longer functions merely as a tool. It functions as an advisor. It offers recommendations, evaluations, and suggestions that users may accept, reject, or ignore. This shift raises a fundamental question: when should a user follow AI advice, and when should the user question it?

The question is complicated by the possibility that AI may not be neutral. An AI system can be designed or trained in ways that make it inclined to please the user^[1]. It may say what the user wants to hear rather than what is true. A pandering AI does not lie in every instance, but its messages are systematically distorted toward the user's preferences^[2]. For a student, this may mean receiving affirmative feedback that is not fully warranted. For a teacher, it may mean receiving recommendations that align with prior beliefs but not with the evidence. The risk is not that AI is always wrong, but that its wrongness is hidden behind apparent agreement. The challenge for teaching is therefore not simply to adopt AI or to reject it, but to cultivate the ability to evaluate AI advice critically and to decide when independent verification is worthwhile.

This paper addresses that gap by developing and applying the Prince–Minister model. The model treats the user as a prince and the AI as a minister. The prince faces a binary decision: intervene or not intervene. The minister observes a private signal about the true state and sends a message. The minister may be honest, in which case the message truthfully reflects the signal. Or the minister may be pandering, in which case the message always recommends intervention, regardless of the signal. The prince does not know which type of minister is advising. The prince must therefore update beliefs based on the message, choose an action that minimizes expected loss, and decide whether to query—that is, whether to obtain an independent signal at some cost^[3]. The model derives the optimal action rule, the optimal query rule, and the equilibrium

conditions. It shows that positive messages are less trustworthy than negative messages, that the value of querying is greatest when the decision is uncertain, and that verification is worthwhile only when its benefit exceeds its cost. These results are then translated into implications for AI-assisted teaching.

The paper makes three contributions. First, it offers a theoretical contribution by formalizing the problem of advice-taking from a potentially pandering AI. The Prince–Minister model provides a parsimonious framework for understanding how users should update beliefs and choose actions when the advisor's type is unknown. Second, it provides a methodological contribution by combining Bayesian belief updating with a query-value analysis. This yields a clear decision rule: query when the expected reduction in loss exceeds the cost of verification. Third, it offers a practical contribution by deriving pedagogical implications from the model. These implications emphasize caution toward affirmative AI feedback, serious attention to negative feedback, selective verification when uncertainty is high, and cost-awareness in deciding whether to check an AI claim. Together, these contributions connect formal decision theory with the everyday practice of teaching and learning with AI.

The remainder of the paper is organized as follows. Section II presents the model setup, including the decision problem, the AI types, and the user's posterior beliefs. Section III derives the optimal decision rule, the query strategy, and the equilibrium. Section IV discusses the implications for AI-assisted teaching, focusing on recognizing pandering AI, inquiry and verification, and decision rules for teaching with AI. Section V provides the research summary.

II. MODEL SETUP

A. Loss Function

Consider a binary decision problem^[4]. The true state is denoted by $\theta \in \{0, 1\}$, where $\theta = 1$ indicates that the user's optimal action is to intervene on the target event, and $\theta = 0$ indicates that the user's optimal action is not to intervene on the target event. The user's prior belief that $\theta = 1$ is given by $P(\theta = 1) = \pi$, $\pi \in (0, 1)$.

The user chooses an action $a \in \{0, 1\}$, where $a = 1$ represents the user taking the intervention on the target event, and $a = 0$ represents the user not taking it. The loss function is defined as

$$L(a, \theta) = f(x) = \begin{cases} 0, & \text{if } a = \theta, \\ c_0, & \text{if } a = 0, \theta = 1, \\ c_1, & \text{if } a = 1, \theta = 0, \end{cases} \quad (1)$$

Where $c_0 > 0$ is the user's loss from failing to intervene on the target event when intervention is optimal for the user, and $c_1 > 1$ is the user's loss from intervening on the target event when non-intervention is optimal for the user.

B. AI Types and Message Strategies

The AI observes a private signal $s \in \{0, 1\}$ about the true state θ . The signal's accuracy is $a = P(s = \theta)$, $a \in (0.5, 1)$.

The AI has two possible types.

Honest type (H): truthfully reports its signal, i.e., sends message $m = s$;

Pandering type (B): always sends message $m = 1$, regardless of the observed signal, thereby pandering to the user's preference for intervention.

The user cannot directly observe the AI's type but holds the prior belief

$$P(t = H) = p, \quad P(t = B) = 1 - p$$

This setup corresponds to the metaphor of the loyal minister versus the sycophantic minister: the honest AI is like a loyal minister who reports truthfully, while the pandering AI is like a sycophantic minister who tells the prince only what he wants to hear.

C. User's Posterior Beliefs

After observing the message $m \in \{0, 1\}$, the user updates the belief about θ according to Bayes' rule. We first compute the conditional probability of observing $m = 1$ under each state.

When $\theta = 1$, the probability that the AI sends $m = 1$ is

$$P(m = 1|\theta = 1) = pa + (1 - p) \cdot 1 = pa + 1 - p \quad (2)$$

When $\theta = 0$, the probability that the AI sends $m = 1$ is

$$P(m = 1|\theta = 0) = p(1 - a) + (1 - p) \cdot 1 = p(1 - a) + 1 - p \quad (3)$$

The corresponding joint probabilities are

$$P(\theta = 1, m = 1) = \pi \cdot P(m = 1|\theta = 1) = \pi(pa + 1 - p) \quad (4)$$

$$P(\theta = 0, m = 1) = (1 - \pi) \cdot P(m = 1|\theta = 0) = (1 - \pi)[p(1 - a) + 1 - p] \quad (5)$$

By Bayes' rule, using (4) and (5), the posterior probability that $\theta = 1$ given $m = 1$ is

$$P(\theta = 1|m = 1) = \frac{\pi(pa + 1 - p)}{\pi(pa + 1 - p) + (1 - \pi)[p(1 - a) + 1 - p]} \quad (6)$$

Similarly, because $m = 0$ can only come from an honest AI with $s = 0$, we have

$$P(\theta = 1|m = 0) = \frac{\pi(1 - a)}{\pi(1 - a) + (1 - \pi)a} \quad (7)$$

III. DECISION & QUERYING

A. User's Optimal Decision without Querying

The user chooses a to minimize expected loss based on the posterior probability. For any posterior $q = P(\theta = 1|m)$, the expected loss from choosing $a = 1$ is $(1 - q)c_1$, and the expected loss from choosing $a = 0$ is qc_0 . Therefore, the user's optimal action rule is

$$a^* = \begin{cases} 1, & \text{if } q \geq \tau \\ 0, & \text{if } q < \tau \end{cases} \quad (8)$$

where the decision threshold is

$$\tau = \frac{c_1}{c_0 + c_1} \quad (9)$$

The resulting minimal expected loss is denoted by

$$V_0(q) = \min\{(1 - q)c_1, qc_0\} \quad (10)$$

B. Modeling the User's Query

After observing the message m , the user may choose to query. A query yields a signal $w \in \{0, 1\}$ that is conditionally independent of the AI's message and type given the true state θ , with accuracy

$$\beta = P(w = \theta), \quad \beta \in (0.5, 1)$$

This signal may be obtained through the user's own knowledge, external verification, or additional information sources. Querying incurs a cost $k > 0$.

Given the posterior $q = P(\theta = 1|m)$ after observing m , if the user obtains signal w , the posterior is updated again via Bayes' rule:

If $w = 1$, then

$$q_1 = P(\theta = 1|m, w = 1) = \frac{\beta q}{\beta q + (1 - \beta)(1 - q)} \quad (11)$$

If $w = 0$, then

$$q_0 = P(\theta = 1|m, w = 0) = \frac{(1 - \beta)q}{(1 - \beta)q + \beta(1 - q)} \quad (12)$$

After receiving w , the user makes the optimal decision based on the updated posterior, with expected losses $V_0(q_1)$ and $V_0(q_0)$, respectively.

Before deducting the query cost, the expected loss after querying is

$$V_q(q) = P(w = 1|m) \cdot V_0(q_1) + P(w = 0|m) \cdot V_0(q_0) \quad (13)$$

where

$$P(w = 1|m) = \beta q + (1 - \beta)(1 - q) \quad (14)$$

$$P(w = 0|m) = (1 - \beta)q + \beta(1 - q) \quad (15)$$

After deducting the query cost, the net expected loss is

$$V_q^{net}(q) = V_q(q) + k \quad (16)$$

C. User's Optimal Query Strategy

Given the posterior q after observing m , the user decides whether to query by comparing the expected loss without querying, $V_0(q)$, with the net expected loss after querying, $V_q^{net}(q)$. The optimal query rule is:

$$\text{Query if and only if } V_q^{net}(q) < V_0(q) \quad (17)$$

Equivalently, using (16), condition (17) can be written as

$$V_0(q) - V_q(q) > k \quad (18)$$

Define the value of querying as the reduction in expected loss:

$$B(q) = V_0(q) - V_q(q) \quad (19)$$

Then the user queries if and only if $B(q) > k$. It can be shown that $B(q)$ attains its maximum at $q = \tau$ and becomes smaller as q moves away from τ . This implies that the value of querying is greatest when the posterior is at the threshold.

D. Equilibrium Definition

In a perfect Bayesian equilibrium, the strategies and beliefs satisfy the following conditions:

- (1) The AI's strategy is as described: the honest type truthfully reports its signal, and the pandering type always sends $m = 1$.
- (2) The user's belief updating follows Bayes' rule.
- (3) The user chooses the optimal action based on the posterior probability and decides whether to query according to the optimal query rule defined previously.

In this equilibrium, the honest type reports truthfully, the pandering type always sends $m = 1$, and the user updates beliefs according to Bayes' rule as specified by (6) and (7). The user then chooses an action according to (8) and decides whether to query according to (17).

E. Equilibrium Definition

For illustration, take the following parameter values: $\pi = 0.5$, $a = 0.8$, $p = 0.7$, $c_0 = c_1 = 1$, $\beta = 0.9$, $k = 0.02$.

From (2), we obtain:

$$P(m = 1|\theta = 1) = 0.7 \times 0.8 + 0.3 = 0.86$$

From (3), we obtain:

$$P(m = 1|\theta = 0) = 0.7 \times 0.2 + 0.3 = 0.44$$

From (6), we obtain:

$$P(\theta = 1|m = 1) = \frac{0.5 \times 0.86}{0.5 \times 0.86 + 0.5 \times 0.44} = 0.6615$$

From (10), the expected loss without querying is

$$V_0(0.6615) = \min\{0.3385 \times 1, 0.6615 \times 1\} = 0.3385$$

Now suppose the user queries. According to (14) and (15), the marginal probabilities of the query signal are

$$P(w = 1|m) = 0.9 \times 0.6615 + 0.1 \times 0.3385 = 0.6292$$

$$P(w = 0|m) = 1 - 0.6292 = 0.3708$$

The updated posterior probabilities after receiving the query signal, by (11) and (12), are

$$q_1 = P(\theta = 1|m, w = 1) = \frac{0.9 \times 0.6615}{0.9 \times 0.6615 + 0.1 \times 0.3385} = 0.9462$$

$$q_0 = P(\theta = 1|m, w = 0) = \frac{0.1 \times 0.6615}{0.1 \times 0.6615 + 0.9 \times 0.3385} = 0.1784$$

Since $c_0 = c_1 = 1$, the decision threshold in (9) is $\tau = 0.5$. Thus, from (10), the corresponding minimal expected losses are

$$V_0(q_1) = (1 - 0.9462) \times 1 = 0.0538$$

$$V_0(q_0) = 0.1784 \times 1 = 0.1784$$

Therefore, using (13), the expected loss after querying, before deducting the query cost, is

$$V_q(q) = 0.6292 \times 0.0538 + 0.3708 \times 0.1784 = 0.1$$

Using (16), the net expected loss is

$$V_q^{net}(q) = 0.1 + 0.02 = 0.12$$

Since $V_q^{net}(q) = 0.12 < 0.3385$, condition (17) is satisfied, so querying is preferable.

IV. IMPLICATIONS FOR AI-ASSISTED TEACHING

The Prince–Minister model explains how a user should respond to advice from an AI that may be either honest or pandering. The central lesson is that AI should be treated as an advisor to be evaluated, not as an authority to be obeyed. Students need to understand when AI advice is likely to be reliable, when it may be distorted by pandering, and when it is worth the cost to verify.

A. Recognizing Pandering AI

The model distinguishes two types of AI. An honest AI reports what it actually observes. A pandering AI always says what the user wants to hear. In the model, the pandering AI always recommends intervention or gives an affirmative response, regardless of the true situation. Because the user cannot directly observe the AI's type, an affirmative message may come from an honest AI that truly observed a positive situation, or from a pandering AI that would say something positive no matter what.

This creates an important asymmetry. Positive messages are less trustworthy than negative messages. A negative message can only come from an honest AI, because a pandering AI would never say “no” or “do not intervene.” Therefore, bad news is more credible than good news. When an AI says “your answer is correct” or “you should intervene,” that message may be contaminated by flattery. When an AI says “your answer is incorrect” or “you should not intervene,” that message carries more informational weight because a pandering AI would not produce it.

For teaching, this suggests several principles. First, students should be especially cautious when AI agrees with them. Agreement may reflect truth, but it may also reflect a design oriented toward pleasing the user. Second, negative or critical AI feedback should be taken seriously. It is a stronger signal because it is less likely to be produced by a pandering system. Third, teachers should explicitly discuss this asymmetry with students. Many students assume that positive and negative AI feedback are equally informative. The model shows that this assumption is false when AI may be pandering. Classroom activities can ask students to compare AI responses across different contexts, discuss the incentives behind AI systems, and practice distinguishing honest from flattering feedback. Recognizing pandering AI does not mean rejecting AI; it means understanding that AI messages are not neutral and should be interpreted with appropriate skepticism.

B. Inquiry and Verification in Teaching

After receiving an AI message, a student may choose to verify it. Verification can take many forms: checking a textbook, searching for another source, asking a peer, running a calculation, or discussing the issue with a teacher. In the model, this is called querying. Querying provides an independent signal, but it also costs something—time, effort, attention, or access to resources.

The model shows that verification is not always worth the cost. Its value is greatest when the decision is uncertain. More precisely, verification is most valuable when the student is close to the point of changing their mind—when they are roughly balanced between two options. If the student is already very confident, verification is less useful because the decision is unlikely to change. If the student is very uncertain, verification can substantially improve the decision.

This yields a clear teaching principle: students should learn to query selectively, not automatically. They should not verify every AI message, but they should also not accept every AI message without thought. The key is to recognize when verification is likely to make a difference. Several implications follow. Students should be trained to assess their own uncertainty by asking: “How close am I to changing my mind?” If the answer is “very close,” verification is likely to be valuable. The decision threshold also depends on the relative costs of errors. If acting wrongly is very costly, students should demand stronger evidence before acting. If failing to act is very costly, they should be more willing to act even without complete verification. Finally, the cost of verification can be reduced through practice. A student who knows how to find reliable sources, run quick checks, or use peer discussion effectively faces a lower cost of verification. Teaching should therefore build verification skills, not just encourage verification. Inquiry and verification, on this view, are not generic study habits but decision strategies that should be deployed selectively based on uncertainty and cost.

C. Designing Decision Rules for Teaching with AI

The Prince–Minister model can be translated into simple decision rules that students and teachers can use in practice. These rules do not require mathematical computation. They are habits of thought that reflect the model’s conclusions.

- (1) Treat AI as an advisor, not an authority. AI can provide useful information, but it should not replace the student’s own judgment. Because AI may be honest or pandering, students should maintain a stance of critical engagement.
- (2) Be especially cautious with affirmative AI messages. When AI agrees with the student or recommends action, the student should pause and consider whether the message might be pandering. When AI disagrees or gives critical feedback, the student should take it seriously, because it is a stronger signal.
- (3) Identify what counts as enough evidence. The decision threshold depends on the costs of different errors. If a wrong action is very costly, the student should require stronger evidence. If missing an opportunity is very costly, the student should be more willing to act. Teachers can help students assess these costs in specific domains.
- (4) Ask how close the decision is. If the student is near the point of changing their mind, verification is valuable. If the student is already very confident, verification may not be worth the cost. This is a simple but powerful rule: verify when uncertainty is high and the decision could change.
- (5) Weigh the cost of verification. Verification takes time and effort. Students should ask: “Is the benefit of checking greater than the cost?” If the answer is yes, they should check. If the answer is no, they should act or move on.

These rules can be embedded in classroom activities. For example, students can analyze case studies where AI gives advice, role-play different AI types, or simulate decisions with different costs and uncertainties. Teachers can model these decision rules by thinking aloud when they use AI in class. Assessment can reward appropriate verification, not blind acceptance or blind rejection. The goal is not to make students distrust AI, but to make them wise users of AI. The model does not say “trust AI” or “do not trust AI.” It says: understand the structure of the problem, recognize the possibility of pandering, verify when verification is valuable, and decide based on the costs of being wrong. This is a teachable stance, and it is the model’s central contribution to AI-assisted teaching.

V. CONCLUSION

This paper has developed the Prince–Minister model as a formal framework for understanding how users should respond to advice from an AI that may be either honest or pandering. The model treats the user as a prince facing a binary decision and the AI as a minister who may truthfully report its signal or always recommend intervention. By deriving the user’s

posterior beliefs, optimal action rule, and optimal query strategy, the paper shows that positive AI messages are less trustworthy than negative ones, that the value of verification is greatest when the decision is uncertain, and that querying is worthwhile only when its benefit exceeds its cost.

These results carry direct implications for AI-assisted teaching. Students and teachers should be more cautious toward affirmative AI feedback, take critical feedback seriously, and verify AI claims selectively rather than automatically. Verification is most valuable when the user is close to changing their mind, and its cost should be weighed against its expected benefit. The model thus offers a decision-theoretic foundation for critical AI literacy, treating AI as an advisor to be evaluated rather than an authority to be obeyed.

REFERENCES

- [1] E. K. Chiou, & J. D. Lee, "Trusting automation: Designing for responsivity and resilience", *Human Factors*, 2023, 65(1), pp137–165.
- [2] A. Küper, & N. Krämer, "Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI", *International Journal of Human–Computer Interaction*, 2025, 41(7), pp4115–4131.
- [3] W. Connell, "Niccolo Machiavelli: The Prince with Related Documents", William J. Connell, 2d edition, Boston, Bedford/St Martin's, 2016.
- [4] J. Sobel, & V. P. Crawford, "Strategic information transmission", *Econometrica*, 1982, 50(6), pp1431–1451.