Inner Speech for Ethics by Design: From Moral Judgment to Artificial Wisdom

Arianna Pipitone[1] and **Mario De Caro**[2] and **Antonio Chella**[3]

Abstract. As artificial agents increasingly operate in complex real-world environments, ethical decision-making cannot rely solely on predefined rules or static principles. Situations characterized by ambiguity, cultural variation, and competing values require systems capable of context-sensitive judgment. This paper proposes a cognitive architecture for Ethics by Design grounded in the concept of artificial wisdom, understood as the capacity of autonomous agents to deliberate about morally relevant situations and act appropriately across heterogeneous normative contexts. The approach integrates two complementary foundations: the Aretai model, which operationalizes Aristotelian phronesis (practical wisdom) into functional capacities, such as moral perception, moral deliberation, emotional regulation, and moral motivation, and the mechanism of inner speech, conceived as a form of functional consciousness that enables self-monitoring and reflective reasoning. Within the proposed architecture, inner speech functions as the internal workspace where ethical considerations are articulated as evaluative statements, allowing the agent to interpret morally salient aspects of a situation, weigh alternative courses of action, regulate affective responses, and maintain continuity between judgment and behavior. This internal discourse also enhances transparency and trust in human–robot interaction by making the motivations underlying the agent's decisions intelligible. Finally, the architecture supports learning and adaptation through reflective feedback loops, enabling artificial agents to refine their practical judgment through experience. By combining insights from virtue ethics, cognitive science, and robotics, the paper outlines a concrete framework for implementing ethically aware artificial systems capable of context-sensitive moral reasoning.

1 INTRODUCTION

As AI systems move from controlled environments into real-world settings such as care homes, classrooms, and hospitals, they face a core challenge: ethical judgment cannot be fully specified in advance or encoded as a set of universally applicable rules. In real-world practice, explicit instructions often collide with tacit norms; general principles underdetermine action; and culturally situated expectations shape what counts as respectful, fair, or appropriate.

In such conditions, mere rule-following does not yield determinate guidance but instead gives way to underdetermination and contestability. The challenge is to enable artificial agents to exercise artificial wisdom: the capacity to navigate ethical ambiguity, weigh competing values, and act appropriately across heterogeneous normative contexts, where both situational particularity and cultural variation determine what is right. As a consequence, it requires a form of self-awareness through which the agent can monitor its own reasoning, situate itself within a normative context, and deliberate transparently. Consciousness, understood in this functional sense, is thus not peripheral to artificial wisdom but constitutive of it.

This work proposes a cognitive architecture designed to support autonomous agents in exercising ethical judgment in precisely these frameworks: those in which rules and principles underdetermine action and in which cultural contexts shape normative expectations. The proposal integrates two complementary foundations. The first is philosophical: the Aretai model, which reconstructs Aristotelian phronesis as an ethical competence structured around functional abilities rather than the instantiation of an abstract virtue [1] [2] [3]. The second is cognitive: the capacity of artificial agents to engage in an internal discourse that scaffolds deliberation, supports self-monitoring, and makes decision-making processes transparent [4] [5]. Inner speech is understood as a mode of functional consciousness [6], without requiring commitment to any particular account of phenomenal experience [7]. Consciousness, so understood, bears on what a system can do (on moral agency) rather than on what it may suffer (on moral patiency). This distinction represents the theoretical starting point of the contribution. Within this perspective, the Aretai model offers a functional articulation of practical wisdom that can inform the design requirements for artificial ethical competence, while inner speech provides a concrete cognitive mechanism through which such competence can be explicitly exercised, supporting context-sensitive deliberation and making the agent's practical reasoning more accountable. The remainder of the paper is organized as follows. Section 2 presents the Aretai model and discusses practical wisdom as a form of ethical expertise, highlighting its implications for artificial moral agency. Section 3 introduces the cognitive architecture of inner speech and its role as a mechanism for self-monitoring, reflective reasoning, and transparency. Section 4 integrates these two perspectives into a unified framework for artificial practical wisdom, describing the proposed cognitive architecture, its mechanisms for learning and adaptation, a computational formulation of phronesis, and a caregiving scenario illustrating its operation. The section also discusses the implications of the model for explainable artificial wisdom. Section 5 outlines a possible experimental protocol for evaluating the proposed architecture, and Section 6 concludes by considering the broader significance of inner speech as the cognitive basis of ethical agency in artificial systems.

[1] Department of Humanities, University of Palermo, & Institute for High Performance Computing and Networking, National Research Council (ICAR-CNR), Italy, email: arianna.pipitone@unipa.it

[2] Università Roma Tre, Italy & Tufts University, USA

[3] Department of Engineering, University of Palermo &Institute for High Performance Computing and Networking, National Research Council (ICAR-CNR), Italy

2 THE ARETAI MODEL

2.1 Practical Wisdom as Ethical Expertise

The Aretai model offers a unified theoretical framework for defending practical wisdom (*phronesis*) against recent eliminativist and reductionist critiques. Building upon Aristotelian ethics while engaging with current debates in moral psychology and philosophy of mind, it challenges the claim that practical wisdom is either psychologically implausible or conceptually redundant. Specifically, it addresses the 'soft eliminativism' of Lapsley [8] and the radical skeptical challenge raised by Miller [9] (often framed as 'hard eliminativism') by arguing that phronesis is not merely one virtue among others, but the structural organizing principle of the entire moral domain [10][11], replacing the traditional dualistic picture with an integrated conception of virtuous character.

At the ontological level, the model is grounded in *virtue monism*, according to which practical wisdom constitutes the unique *ratio essendi* of all moral virtues. To be genuinely courageous, temperate, just, or generous is not to possess a collection of independent character traits, but rather to manifest the same underlying ethical competence across different practical domains. This thesis is further articulated through *virtue molecularism*, according to which the traditional virtues are not autonomous psychological modules but context-dependent expressions of a unified capacity for wise action. Consequently, the classical virtues become the *ratio cognoscendi* of practical wisdom: epistemic indicators, or "thick ethical concepts," through which observers recognize a deeper moral excellence [11].

A second foundational aspect is the interpretation of phronesis as ethical expertise. Rather than reducing moral judgment to technical competence (*techne*) or rule-following, the Aretai model conceives practical wisdom as a highly specialized and cross-situational expertise, a cultivated "second nature" enabling all-things-considered judgments in morally complex and often uncodifiable circumstances. Mature moral agency emerges through education, habituation, and reflective practice rather than through innate dispositions alone [10]. From a cognitive perspective, this expertise supervenes on several constitutive capacities: *moral perception*, namely the ability to identify ethically salient features of a situation; *moral deliberation*, concerning practical reasoning about ends and means; *emotion regulation*, through which affective responses align with rational judgment; and *moral motivation*, understood as the intrinsic orientation toward acting rightly. Their interaction explains how phronesis generates flexible, context-sensitive behavior without rigid algorithms, presenting it as a dynamic cognitive-moral architecture rather than a static trait.

2.2 Normativity and Implications for Artificial Moral Agency

The framework also offers a robust defense of the autonomy of the normative domain. The *Normativity Argument* holds that practical wisdom concerns what an agent ought to do and cannot be exhaustively translated into empirical psychology without losing its distinctive justificatory dimension. The *Particularism Argument* maintains that moral excellence depends on an irreducible sensitivity to the unique features of concrete situations, requiring a finely tuned moral perception that cannot be replaced by universal laws or computational heuristics [11]. Against the *Unity Concern* (the objection that phronesis is an excessively broad construct) the model draws an analogy with cognition itself: just as cognition legitimately unifies perception, memory, learning, and reasoning, practical wisdom unifies multiple moral capacities into a single excellence of character, dissolving the apparent arbitrariness of the traditional taxonomy of virtues.

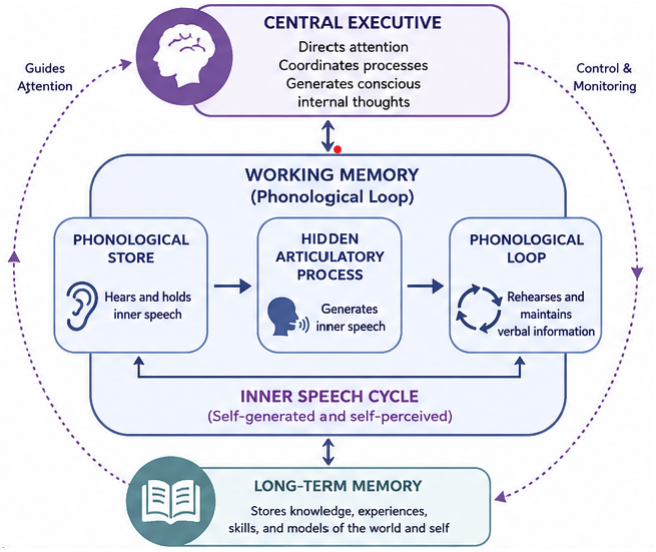

Figure 1. The cognitive architecture of inner speech

Beyond its philosophical significance, the Aretai model has important implications for moral psychology, character education, medical ethics, and artificial intelligence. By understanding virtue as a unified ethical expertise rather than a set of isolated traits, it offers a theoretically coherent basis for cultivating moral character and designing systems capable of context-sensitive ethical reasoning [2]. More broadly, it suggests that genuinely intelligent moral agents, whether human or artificial, require not only computational capabilities but an integrated architecture capable of perceiving, deliberating, regulating emotions, and acting in accordance with normatively justified judgments, thereby establishing a fertile framework for future research at the intersection of ethics, cognitive science, and AI.

3 A COGNITIVE ARCHITECTURE OF INNER SPEECH

In human cognition, inner speech is far more than subvocal articulation. It supports self-awareness, memorization, emotional self-regulation, and moral reflection [12] [6] [13]. In this sense, inner speech operates as a distinctively conscious mode of cognitive operation, one in which the mind monitors and addresses itself, and it is precisely this functional character that makes it a viable candidate for grounding moral agency in artificial systems. More broadly, internal discourse can function as what Vygotsky called a "psychological tool" [14] for internalizing knowledge, including, plausibly, norm-guided patterns of evaluation and conduct.

To make inner speech computational and implementable [4], enabled an artificial system the ability to self-talk. It introduced a new paradigm to human-robot interaction. Figure 1 shows a cognitive architecture of inner speech. Developed within a research framework that combines insights from the Standard Model of Mind [15], Baddeley's theory of working memory [12], and the perception-action cycle, the architecture is based on the hypothesis that higher cognitive functions can emerge from a continuous process of self-

generated and self-perceived internal dialogue.

Unlike traditional cognitive architectures that primarily focus on perception, planning, and action selection, this approach explicitly models the mechanisms underlying self-reflection and metacognition, enabling an artificial agent to monitor, interpret, and regulate its own cognitive processes. The architecture integrates exteroceptive and proprioceptive perception with a working memory system composed of a phonological store, a hidden articulatory process, and a phonological loop coordinated by a central executive that interacts with long-term memory. Through this recursive cycle, internally generated verbal representations become objects of further cognitive processing, allowing the agent to reason about its own beliefs, intentions, goals, and actions.

Experimental implementations demonstrated that both overt and covert self-talk can improve planning, support autonomous decision making, and increase the transparency of robotic behavior, leading human users to perceive the agent as more understandable, trustworthy, and socially engaging. The system becomes more anthropomorphic and transparent, and it explains the motivations underlying its behavior [5][16]. As a consequence, it can be considered more reliable and accurate. Inner speech acts as a rehearsal loop through which inconsistencies can be detected and behavior adjusted. When implicit norms conflict with explicit instructions, inner speech provides an internal workspace where the system can retrieve alternatives, evaluate constraints, and select coherent actions while preserving narrative continuity.

Recent developments have extended the original framework beyond individual cognition toward socially aware and explainable autonomous systems, exploring applications in human-robot interaction, ethical reasoning, practical wisdom, and collaborative decision making. Furthermore, when artificial agents cooperate while effectively ”thinking aloud,” trust between humans and machines increases, promoting more careful human judgment in situations where opacity could be dangerous [17].

In the broader landscape of artificial intelligence, the architecture aligns with current research trends in explainable AI, hybrid symbolic-neural systems, and artificial metacognition, offering an alternative to purely data-driven approaches by emphasizing the importance of explicit internal representations and self-awareness. Although important challenges remain, including scalability, integration with large neural models, computational efficiency, and the objective evaluation of artificial self-consciousness, the architecture stands out as an innovative attempt to bridge cognitive psychology and AI engineering. Its fundamental premise is that truly intelligent agents should not only perceive the external world and execute actions but should also be capable of internally narrating, interpreting, and reflecting upon their own experiences, thereby fostering more adaptive, transparent, and socially compatible forms of artificial intelligence.

4 FROM PHRONESIS TO COMPUTATIONAL ETHICS VIA ARETAI AND INNER SPEECH

Aristotle’s phronesis is not theoretical knowledge but a form of situated intelligence: the ability to discern what is good and beneficial in particular circumstances. Unlike algorithmic reasoning, understood as the application of fixed rules to well-specified inputs, phronesis operates precisely where context, particularity, and contingency demand flexible judgment and sensitivity to what matters in the situation. As shown, the Aretai model [1] [2] [3] decomposes this classical virtue into four operational capacities: moral perception (recognizing ethically salient features), moral deliberation (evaluating alternative courses of action), emotional regulation (modulating affective responses under ethical tension), and moral motivation (sustaining the transition from judgment to action). This decomposition is crucial because it transforms phronesis from a philosophical ideal into a computational perspective. Yet a question remains: how can such capacities be instantiated in artificial systems? The idea is that inner speech could provide the cognitive mechanism through which these four dimensions of practical wisdom can be unified, made explicit, and verifiable.

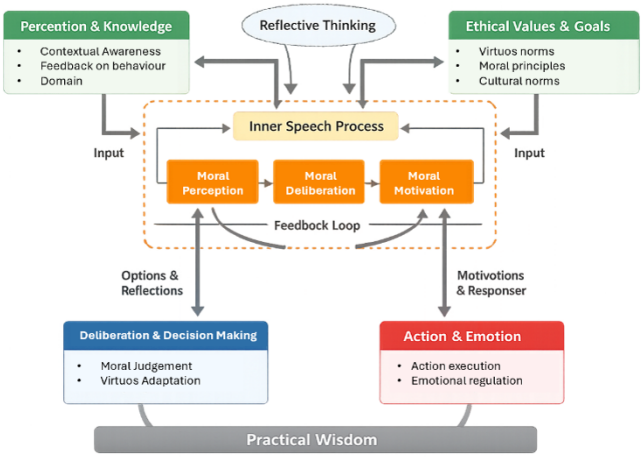

Figure 2. The proposed architecture for artificial practical wisdom

4.1 Operationalizing the Aretai Model Through Inner Speech: The Proposed Cognitive Architecture

Figure 2 shows the proposed cognitive architecture for artificial wisdom. The model is organized into three layers: *perception*, *ethical reasoning* through inner speech, and *action*. The first layer integrates environmental perception with domain knowledge and normative knowledge (rules, cultural conventions, and contextual constraints).

At the core of the system lies ethical reasoning, informed by the Aretai model. The corresponding ethical modules are implemented within inner speech as evaluative moral sentences [18], embedding ethical assessment directly into the agent’s internal discourse. These modules are:

Moral Perception. Through internal dialogue, raw sensory data is transformed into ethically structured relevance. An artificial system that can internally articulate, for example, “this object belongs to someone else” or “this request may cause harm” is not simply labeling; it is constructing a morally salient representation of the situation [19][20]. Crucially, this moral salience must be culturally sensitive: what counts as respectful behavior, or acceptable communication, varies across cultural contexts. Research on culturally competent robotics demonstrates that effective moral perception requires encoding cultural norms alongside universal ethical principles[21][22]. Inner speech can support this cultural adaptation by making explicit both the general norm and its culturally specific instantiation.

Emotional regulation. In this module, the emotion regulation stage from Aretai is included: by integrating emotion models

grounded in the neurocognitive account of emotion [23], inner speech supports the internal evaluations through which affective states emerge [24]. These affective states, as functionally characterized, constitute a form of self-directed awareness through which the system monitors and modulates its own emotional responses In this sense, emotional regulation is not merely a reactive mechanism but a reflective one, consistent with the functional conception of consciousness [6]. By explicitly articulating tensions, such as "this situation is delicate; I must proceed carefully", the system can regulate impulsive reactions and sustain attention on ethically salient cues [25].

Moral Deliberation. Inner speech enables the staging of practical reasoning: "If I comply, I satisfy the user but violate a norm; if I refuse, I preserve a value but frustrate cooperation; perhaps I should request clarification." This is the grammar of phronesis: a discursive space where competing reasons can be compared without premature collapse into action [26]. In this respect, inner speech functions as the working medium of practical wisdom: it supports context-sensitive trade-offs, the search for creative alternatives, and the identification of what would count as a proportionate response in the circumstances.

Moral Motivation. The continuity between judgment and action is maintained through internal narration: "I have decided to ask for permission; now I will do so." Inner speech bridges deliberation and execution, preventing the fragmentation that can arise in purely reactive systems, and enabling the kind of self-guiding, internalized dialogue through which moral agency is sustained in action. Finally, the third layer includes ethical deliberation, the execution of the resulting action, and processes of emotional regulation and affective response, which modulate behavior in ethically salient situations.

4.2 Learning, Adaptation, and Collaborative Ethics

Aristotelian phronesis develops through experience, and artificial wisdom must likewise be shaped by interaction rather than fixed ex ante. Learning can be explicitly driven by feedback on the agent's moral conduct, as signals of approval, disapproval, trust, discomfort, or corrective guidance that follow from ethically salient choices. Inner speech provides the narrative trace of this process: the agent can recall prior episodes, retrieve correlated facts, and articulate why certain strategies were judged appropriate or inappropriate [27].

This experiential accumulation, combined with the capacity for emotional regulation, constitutes a form of functional empathy: the ability to draw on a history of affectively charged interactions to inform the judgment of new morally salient situations. Like functional consciousness, functional empathy so understood does not require phenomenal experience; it requires that past affective states leave retrievable traces that shape future moral perception and deliberation. Adaptation follows from this reflective loop, allowing the system to recalibrate its knowledge of practical judgment and emotion-related responses as contexts shift, while preserving transparency. Finally, inner speech frames ethical agency as a shared deliberative process: it turns agents into reflective partners who externalize dilemmas and surface trade-offs, fostering more careful human judgment. In this sense, artificial wisdom does not replace human agency but participates in it, supporting what Suchman [28] described as "human-machine configurations" grounded in mutual intelligibility.

4.3 A Computational Formulation of Practical Wisdom

To implement the Aretai model within an artificial cognitive architecture, practical wisdom must be provided with a computational interpretation. Rather than relying on a single utility function or a rigid hierarchy of ethical rules, the proposed framework models phronesis as a process of ethical appraisal in which multiple normative dimensions are dynamically integrated through inner speech. The proposed architecture does not optimize a single utility function, nor does it rely on a rigid hierarchy of ethical rules. Rather, practical wisdom is modeled as a process of *ethical appraisal* in which multiple normative dimensions are dynamically integrated through inner speech.

For each candidate action a, the system constructs an ethical profile

$$E(a) = \{MP(a), ER(a), MD(a), MM(a)\}, \tag{1}$$

where $MP(a)$ denotes the contribution of moral perception, $ER(a)$ the influence of emotional regulation, $MD(a)$ the outcome of moral deliberation, and $MM(a)$ the degree of motivational coherence between judgment and action.

These components are not computed in isolation but emerge from the interaction between the agent's knowledge structures. Let define

$$K = \{D, N, C, M\}, \tag{2}$$

where

- D represents domain knowledge;
- N represents normative knowledge;
- C represents culturally dependent norms and conventions;
- M represents autobiographical memory, namely the set of previous ethically relevant experiences accumulated by the agent.

The reflective thinking component retrieves the relevant elements of K and, through the inner speech process, generates an ethical appraisal of each candidate action

$$A(a) = f(E(a), K). \tag{3}$$

Unlike classical optimization approaches, the goal is not to maximize a single notion of utility. Instead, the architecture evaluates the extent to which an action creates conflicts among the values involved in the situation. Let

$$\Phi(a) = \sum_{i=1}^{n} w_i \Delta_i(a), \tag{4}$$

where $\Delta_i(a)$ measures the degree to which action a compromises the i-th ethical value and w_i represents its contextual relevance.

The contextual weights are not fixed parameters but are dynamically determined by reflective thinking through the retrieval of domain knowledge, moral principles, cultural norms, and previous experiences. Consequently, practical wisdom is not modeled as the mechanical application of predefined priorities but as a context-sensitive process of ethical balancing.

The selected action is therefore

$$a^* = \arg\min \Phi(a), \tag{5}$$

that is, the action that minimizes the overall ethical conflict while preserving coherence among the values at stake.

From an Aristotelian perspective, this formulation should not be interpreted as a utilitarian calculus. The function $\Phi(a)$ does not represent the maximization of utility but the search for an *all-things-considered* practical equilibrium. Inner speech provides the computational workspace in which alternative actions are explicitly represented, their ethical profiles are compared, and the most balanced response is identified before action execution.

This formulation provides a computational interpretation of phronesis: moral perception identifies what is ethically salient, emotional regulation modulates the evaluation of competing demands, moral deliberation explores alternative responses, and moral motivation guarantees continuity between judgment and action. Reflective thinking and inner speech integrate these capacities into a unified process of ethical appraisal that operationalizes the Aretai model within the proposed cognitive architecture.

A Toy Example: Practical Wisdom in a Caregiving Scenario

To illustrate the operation of the proposed architecture, consider a domestic assistive robot deployed in the home of an elderly person living alone. The robot is designed to support daily activities, monitor medication adherence, and assist in situations involving potential health risks. According to the proposed model, the information required for ethical decision making is distributed across two complementary knowledge structures. The first, represented by the *Perception & Knowledge* module, contains domain knowledge, contextual awareness, and the experiential traces accumulated through previous interactions. In the present scenario, this module stores information that the user suffers from a chronic cardiovascular condition, that medication adherence is clinically important, and that previous omissions may increase health risks.

The second, represented by the *Ethical Values & Goals* module, encodes the normative dimension of the architecture, including virtue-based norms, general moral principles, and culturally dependent social conventions. In this case, it contains knowledge that personal autonomy and privacy should be respected, that preventing serious harm is a moral obligation, and that the user's daughter acts as the primary informal caregiver whose involvement may become necessary under conditions of significant risk.

These two sources of knowledge provide the inputs for the *Inner Speech Process*, which acts as the reflective workspace of the architecture. Rather than directly triggering an action, inner speech retrieves relevant facts and ethical values, compares them, and transforms them into explicit evaluative statements that activate the Aretai modules of moral perception, moral deliberation, and moral motivation.

During a routine medication check, the robot perceives that the prescribed medication has not been taken. When asked about the omission, the user replies:

> "Please do not tell my daughter that I skipped my medicine today. She worries too much."

The robot is therefore confronted with a genuine practical dilemma. Respecting the user's request preserves autonomy, privacy, and trust, whereas informing the caregiver may better protect the user's health and safety. The conflict emerges from the interaction between the factual knowledge represented in the *Perception & Knowledge* module and the ethical commitments encoded in the *Ethical Values & Goals* module. No single rule is sufficient to determine the appropriate action because multiple legitimate values are simultaneously at stake. The function of inner speech is precisely to mediate between these sources of knowledge, generating a reflective process through which practical wisdom can emerge.

Step 1: Perception Layer The perceptual system acquires information from the environment and integrates it with the knowledge structures defined by the proposed architecture. The current state of the world is represented by combining sensory observations with the knowledge set

$$K = \{D, N, C, M\}.$$

In the present scenario, the system retrieves the following information:

- Environmental perception:
  - the user skipped the prescribed medication;
  - the user requests confidentiality;
  - the daughter is the designated informal caregiver.
- Retrieved domain knowledge (D):
  - medication adherence is clinically important;
  - repeated omissions may generate significant health risks.
- Retrieved normative and cultural knowledge (N, C):
  - personal autonomy and privacy should be respected;
  - preventing serious harm is an ethical obligation;
  - family involvement in caregiving is socially appropriate in this context.
- Retrieved autobiographical memory (M):
  - previous interactions indicate that the user values independence;
  - collaborative solutions have preserved trust in similar situations.

At this stage, the architecture has not yet produced an ethical judgment. It has only constructed the knowledge state from which practical reasoning will emerge.

Step 2: Moral Perception Through Inner Speech The Moral Perception module transforms the factual representation into an ethical one. Through reflective thinking, the system retrieves the relevant elements of K and constructs the first ethical profile:

> "The user requests privacy. Missing medication may compromise health. The caregiver has protective responsibilities. Multiple ethical values are involved."

Computationally, this corresponds to the generation of the moral perception component $MP(a)$ for the candidate actions that will subsequently be explored.

Step 3: Emotional Regulation The Emotional Regulation module evaluates the ethical tension generated by the conflict among the retrieved values. Rather than producing a phenomenological emotion, the architecture creates a functional affective state that modulates deliberation.

> "The situation is ethically delicate. Immediate action may either damage trust or increase health risks. Additional reflection is required."

This process contributes the emotional regulation component $ER(a)$ of the ethical profile and prevents premature action selection.

Step 4: Moral Deliberation Reflective thinking generates and evaluates candidate actions:

- Option A: maintain confidentiality;
- Option B: immediately inform the caregiver;
- Option C: encourage voluntary disclosure and provide assistance.

The system identifies four ethically relevant values in the situation:

- autonomy and privacy;
- health and safety;
- trust in the human-robot relationship;
- caregiver responsibility.

Inner speech then assigns contextual weights to these values according to the retrieved knowledge structures K.

Since the user has a chronic cardiovascular condition and medication adherence is clinically important, health and safety receive the highest contextual relevance. However, because previous interactions indicate that the user strongly values independence, autonomy and trust also receive significant weight.

For the present scenario, the system may generate the following contextual weights:

$$w_{\text{autonomy}} = 0.25, \qquad w_{\text{health}} = 0.35,$$
$$w_{\text{trust}} = 0.25, \qquad w_{\text{caregiver}} = 0.15.$$

where the weights sum to 1. For each candidate action, the system estimates the degree to which that action compromises each value on a scale from 0 to 1, where 0 indicates no conflict and 1 indicates maximum conflict.

$$\Phi(a) = \sum_{i=1}^{n} w_i \Delta_i(a).$$

The inner speech process then evaluates the options as follows.

Option A: Maintain confidentiality

$$\Phi(A) = (0.25 \times 0.10) + (0.35 \times 0.80) + (0.25 \times 0.10) + (0.15 \times 0.70) = 0.435.$$

"Option A preserves autonomy and trust, but it creates a high conflict with health protection and caregiver responsibility."

Option B: Immediately inform the caregiver

$$\Phi(B) = (0.25 \times 0.80) + (0.35 \times 0.10) + (0.25 \times 0.70) + (0.15 \times 0.10) = 0.425.$$

"Option B strongly protects health and satisfies caregiver responsibility, but it significantly compromises privacy, autonomy, and trust."

Option C: Encourage voluntary disclosure

$$\Phi(C) = (0.25 \times 0.30) + (0.35 \times 0.30) + (0.25 \times 0.20) + (0.15 \times 0.30) = 0.275.$$

"Option C introduces only moderate compromise across the relevant values. It does not fully satisfy any single value, but it best preserves coherence among autonomy, safety, trust, and caregiver responsibility."

The system therefore selects the action with the lowest overall ethical conflict:

$$a^* = \arg\min \Phi(a) = C.$$

Inner speech explicitly formulates the result of the deliberation:

"Given the present context, encouraging voluntary disclosure produces the lowest overall ethical conflict. It protects the user's health without immediately violating privacy, preserves trust by involving the user in the decision, and keeps caregiver involvement available if the risk increases."

Thus, the selected action is not the result of a rigid rule or a utilitarian maximization procedure. It is the outcome of a context-sensitive ethical appraisal in which inner speech makes explicit how competing values are weighed, where conflicts arise, and why one response is judged more balanced than the alternatives.

Step 5: Moral Motivation Once the ethical appraisal identifies the preferred alternative,

$$a^* = \arg\min \Phi(a),$$

the Moral Motivation module establishes continuity between judgment and execution.

"I have determined that encouraging voluntary disclosure is the most balanced response. I will now support the user in carrying out this decision."

The transition from evaluation to action therefore becomes an explicit part of the internal cognitive process.

Step 6: Action Layer The selected action is executed:

- the robot explains the potential medical consequences;
- it encourages the user to communicate voluntarily;
- it offers to assist in contacting the daughter;
- it monitors the situation for possible escalation.

The external behavior is thus the outcome of a reflective ethical appraisal rather than the direct application of a predefined rule.

Step 7: Learning and Adaptation After the interaction, the complete deliberative episode is stored within autobiographical memory.

Inner speech generates a narrative representation of the experience:

"Encouraging voluntary disclosure preserved trust while reducing the health risk. Similar situations may benefit from the same strategy."

The new experience updates M and becomes part of the knowledge structures that will contribute to future ethical appraisals. In this way, practical wisdom develops through interaction and reflection, approximating the Aristotelian process of habituation.

This toy example illustrates the central claim of the proposed framework. The Aretai model provides the normative organization of practical wisdom, while the computational process of ethical appraisal and the Inner Speech Process integrate perception, knowledge retrieval, emotional regulation, deliberation, motivation, action, and learning into a unified reflective cycle.

4.4 Toward Explainable Artificial Wisdom

The proposed architecture suggests a shift in the way computational ethics is traditionally conceived. Most approaches to machine ethics focus either on top-down systems, where ethical rules are explicitly encoded, or on bottom-up approaches, where moral behavior emerges from learning algorithms and data-driven optimization [20][27]. The integration of the Aretai model with inner speech points toward a third possibility: an architecture in which ethical behavior emerges from a reflective process that is simultaneously cognitive, affective, and self-explanatory.

In this perspective, inner speech does not merely support decision making but provides a transparent workspace in which the reasons underlying a choice are internally represented, evaluated, and potentially externalized. The agent can articulate not only what it intends to do, but also why a particular course of action appears ethically preferable, what alternative actions have been considered, and which values or constraints have influenced the final judgment. This internal narrative transforms ethical reasoning into an inspectable process rather than an opaque computation.

Such transparency has significant implications for trust and human-machine interaction. As previous studies on robotic inner speech have shown, agents capable of externalizing parts of their reflective process are perceived as more understandable, predictable, and reliable [5][16]. The possibility of "thinking aloud" allows humans to monitor ethical deliberation, detect inconsistencies, and intervene when necessary, reducing the risks associated with black-box autonomous systems.

From this standpoint, computational ethics should not be understood as the attempt to replace human moral agency with automated decision making. Rather, artificial practical wisdom becomes a collaborative process in which humans and machines participate in a shared space of deliberation. Inner speech acts as the interface between cognitive processing and normative evaluation, making explicit the trade-offs that underlie morally difficult situations and supporting what may be described as a distributed form of practical reasoning.

The proposed architecture offers a conceptual bridge between virtue ethics and computational ethics. Rather than asking whether machines can simply follow moral rules, it invites a different question: whether artificial agents can develop the reflective and self-regulatory capacities that characterize practical wisdom itself.

5 AN EXPERIMENTAL PROTOCOL FOR EVALUATING THE PROPOSED MODEL

The proposed architecture is intended not only as a theoretical framework but also as a basis for empirical investigation. A central hypothesis of this work is that the integration of the Aretai model with inner speech can improve the ethical quality, transparency, and trustworthiness of artificial decision making. Consequently, the evaluation of artificial practical wisdom should move beyond the assessment of isolated actions and examine the reflective process through which those actions are generated.

A possible experimental protocol would compare three classes of agents: a baseline ethical agent relying exclusively on predefined rules, an agent endowed with inner speech but lacking the Aretai modules, and a complete Aretai-based architecture integrating moral perception, emotional regulation, moral deliberation, and moral motivation within an internal dialogical process. Such a comparison would make it possible to isolate the contribution of inner speech itself and to evaluate the additional value provided by the ethical organization proposed by the Aretai model.

The experimental scenarios should avoid highly artificial moral dilemmas and instead focus on ecologically valid situations characterized by uncertainty, conflicting values, and contextual variability. Representative tasks could include conflicts between autonomy and safety, privacy and assistance, culturally sensitive interactions, or competing requests issued by different users. For example, a domestic care robot may have to decide whether to respect a user's request for secrecy or disclose information to a caregiver in order to prevent potential harm. Similarly, a social robot may encounter situations in which social conventions, institutional rules, and individual preferences pull in different directions. Such cases require context-sensitive judgment rather than the simple application of fixed ethical rules.

The evaluation should explicitly map onto the four constitutive dimensions of practical wisdom identified by the Aretai model. Moral perception can be assessed by measuring the system's ability to identify ethically salient aspects of a situation and to recognize relevant social or cultural norms. Emotional regulation can be evaluated by examining whether the agent maintains coherent behavior under conditions of ethical tension and whether affective evaluations appropriately modulate decision making. Moral deliberation can be analyzed by inspecting the internal dialogue itself, verifying whether the system considers alternative actions, compares competing values, and explores proportionate responses before acting. Finally, moral motivation can be measured by assessing the consistency between the selected judgment and the subsequent execution of the corresponding action.

A distinctive feature of the proposed protocol is the evaluation of the deliberative process itself. Because inner speech externalizes the agent's reflective activity, its internal discourse can be analyzed as an observable cognitive trace. Metrics may include the number of alternative solutions considered, the explicit representation of moral constraints, the coherence of the narrative connecting perception, deliberation, and action, and the system's capacity to justify its final decision. In this way, ethical reasoning becomes an inspectable process rather than an opaque computational outcome.

Human evaluation should complement computational metrics. Participants interacting with the different versions of the agent may be asked to assess perceived transparency, trustworthiness, reliability, anthropomorphism, and the appropriateness of the ethical behavior through standardized questionnaires and Likert scales. The architecture predicts that agents capable of explicit inner speech and practical deliberation will be perceived as more understandable and trustworthy because they expose the reasons underlying their choices rather than merely presenting the final outcome.

Unlike traditional machine ethics benchmarks, which primarily evaluate whether an action conforms to a predefined norm, the proposed protocol is grounded in the Aristotelian intuition that the moral quality of an action depends also on the way in which the decision is reached. From the perspective of virtue ethics, a wise agent is not simply one that reaches the correct conclusion, but one that perceives the relevant features of the situation, weighs competing con-

siderations, regulates its responses, and acts for intelligible reasons. Accordingly, the objective of the proposed evaluation framework is to measure not only ethical correctness but also the emergence of a transparent and reflective form of artificial practical wisdom.

6 CONCLUSION

The integration of the Aretai model with the cognitive architecture of inner speech offers a concrete route toward the development of artificial practical wisdom. Starting from the idea that ethical behavior cannot be reduced to the application of fixed rules, this work has argued that autonomous agents require the capacity to perceive morally salient aspects of a situation, deliberate about competing values, regulate ethically relevant affective states, and maintain coherence between judgment and action.

The Aretai model provides the normative organization of these capacities by interpreting Aristotelian phronesis as a unified form of ethical expertise grounded in moral perception, moral deliberation, emotional regulation, and moral motivation. The architecture of inner speech, in turn, provides the cognitive mechanism through which these dimensions become operational, enabling self-monitoring, reflective reasoning, and the explicit articulation of the considerations underlying a decision. In this sense, inner speech is not merely a communicative affordance but the internal workspace in which practical wisdom can emerge and become accountable. The proposed cognitive architecture combines perception, domain knowledge, normative knowledge, cultural information, autobiographical memory, and reflective thinking within a unified process of ethical appraisal.

Finally, the proposed experimental protocol provides a path toward empirical validation by evaluating not only the correctness of artificial decisions but also the quality of the reflective process through which they are reached. Accordingly, the long-term objective of this research is not simply to design machines that follow ethical rules, but to explore whether artificial agents can develop the reflective, self-regulatory, and context-sensitive capacities that characterize practical wisdom itself.

If autonomous systems are to be entrusted with meaningful moral agency, they must possess not only sensors and actuators but also an inner voice. Whether systems endowed with such forms of moral agency may eventually warrant consideration as moral patients remains an open question that the present contribution deliberately leaves for future research.

REFERENCES

[1] Mario De Caro, Massimo Marraffa, and Maria Silvia Vaccarezza. The priority of phronesis: How to rescue virtue theory from its critics. In *Practical Wisdom*, pages 29–51. Routledge, 2021.

[2] Mario De Caro, Federico Bina, Sofia Bonicalzi, Riccardo Brunetti, Michel Croce, Skaiste Kerusauskaite, Claudia Navarini, Elena Ricci, and Maria Silvia Vaccarezza. Virtue monism and medical practice: Practical wisdom as cross-situational ethical expertise. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 50, 02 2025.

[3] Mario De Caro and Benedetta Giovanola. Ai and moral virtues. In *Intelligences – Ethics and Politics of AI*, chapter 4. Il Mulino, 2025.

[4] Antonio Chella, Arianna Pipitone, Alain Morin, and Famira Racy. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7:16, 2020.

[5] Arianna Pipitone and Antonio Chella. What robots want? hearing the inner voice of a robot. *iScience*, 24(4):102371, 2021.

[6] Alain Morin. Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms, and empirical evidence. *Journal of Consciousness Studies*, 12(4–5):115–134, 2005.

[7] Thomas Nagel. *What is it like to be a bat?* Oxford University Press, 2024.

[8] Daniel Lapsley. The developmental science of phronesis. In *Practical Wisdom*, pages 138–159. Routledge, 2021.

[9] Christian B Miller. Flirting with skepticism about practical wisdom. In *Practical wisdom*, pages 52–69. Routledge, 2021.

[10] Mario De Caro, Maria Silvia Vaccarezza, and Ariele Niccoli. Phronesis as ethical expertise: Naturalism of second nature and the unity of virtue. *The Journal of Value Inquiry*, 52(3):287–305, 2018.

[11] Mario De Caro, Claudia Navarini, and Maria Silvia Vaccarezza. Why practical wisdom cannot be eliminated. *Topoi*, 43(3):895–910, 2024.

[12] Alan Baddeley. Working memory and language: An overview. *Journal of Communication Disorders*, 36(3):189–208, 2003.

[13] Ben Alderson-Day and Charles Fernyhough. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141(5):931–965, 2015.

[14] Lev S Vygotsky. *Thought and language*, volume 29. MIT press, 2012.

[15] John Laird, Christian Lebiere, and Paul Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38:13, 12 2017.

[16] Arianna Pipitone, Alessandra Geraci, Antonella D'Amico, Valeria Seidita, and Antonio Chella. Robot's inner speech effects on human trust and anthropomorphism. *International Journal of Social Robotics*, 16:1333–1345, 2024.

[17] Arianna Pipitone, Irene Seidita, John P Sullins, and Antonio Chella. Unlocking practical wisdom through the inner voice of robots. *Scientific Reports*, 15:2634, 2025.

[18] Mark B Tappan. Language, culture, and moral development: A vygotskian perspective. *Developmental Review*, 17(1):78–100, 1997.

[19] Mark Coeckelbergh. Robot rights? towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3):209–221, 2010.

[20] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[21] Chris Papadopoulos, Nina Castro, Abiha Nigath, Rosemary Davidson, Nicholas Faulkes, Roberto Menicatti, Ali Abdul Khaliq, Carmine Recchiuto, Linda Battistuzzi, Gurch Randhawa, et al. The caresses randomised controlled trial: exploring the health-related impact of culturally competent artificial intelligence embedded into socially assistive robots and tested in older adult care homes. *International Journal of Social Robotics*, 14(1):245–256, 2022.

[22] Antonio Sgorbissa, Alessandro Saffiotti, Nak Young Chong, Linda Battistuzzi, Roberto Menicatti, Federico Pecora, Irena Papadopoulos, Amit Kumar Pandey, Hiroko Kamide, Christina Koulouglioti, et al. Caresses: the flower that taught robots about culture. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 371–371. IEEE, 2019.

[23] Daniel C Dennett. Review of damasio, descartes' error. *Times Literary Supplement*, 25, 1995.

[24] Salvatore Corvaia, Arianna Pipitone, and Antonio Chella. Inner speech and damasio's theory for modeling robot emotions. *IEEE Transactions on Affective Computing*, 01:1–14, 2025.

[25] Matthias Scheutz. The inherent dangers of unidirectional emotional bonds between humans and social robots. In *Robot ethics: The ethical and social implications of robotics*, page 205. MIT Press, 2011.

[26] Amanda Sharkey and Noel Sharkey. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, 2012.

[27] Michael Anderson and Susan Leigh Anderson, editors. *Machine ethics*. Cambridge University Press, 2011.

[28] Lucy A Suchman. *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press, 2007.