Journal of Distributed Computing and Systems(JDCS)
Vol.9, Issue 1, 2026, Pages 111-126

RAVA: An Integrated Cognitive-Spiritual Architecture for Value-Aligned Decision-Making in Autonomous Intelligent Agents

Arash Khosravi*, Zahra Najafi
Faculty of Engineering, Mahallat Institute of Higher Education, Mahallat, Iran.
**khosravi.280@gmail.com*

Article History:
Received: 20 May 2026
Received in revised form: 30 June 2026
Accepted: 28 August 2026
Available online: 29 September 2026

Abstract

Cognitive architectures are computational frameworks designed to simulate human mental processes such as perception, reasoning, and decision-making. Existing architectures, however, are largely built to optimize computational efficiency and generally overlook the ethical and spiritual dimensions of human thought. This paper introduces RAVA (Revelation-Aware Value Architecture), a novel cognitive-spiritual architecture for intelligent decision-making grounded in Islamic anthropology. Because its five minds operate as coordinating, semi-autonomous units under a shared executive, RAVA is also naturally suited to the design of intelligent agents, the paradigm behind most software built today. Drawing on the model of five parallel minds, Theory of Mind, and Islamic ethical principles, RAVA seeks to establish a dynamic balance between cognitive efficiency and spiritual purpose. Unlike classical cognitive architectures, which treat ethics, if addressed at all, as an external add-on module, RAVA internalizes the spiritual dimension as a pervasive variable embedded within every layer of the system. At the core of this architecture lies the Nafs and Taqwa Center (NTC), a dedicated component that identifies the different states of the self (nafs), evaluates the purity of intention, and provides ethical guidance — transforming the system from a purely computational decision-maker into a moral-formative companion.

The architecture was evaluated through an exploratory qualitative assessment using open, axial, and selective coding procedures derived from Grounded Theory, a systematic method for building theoretical frameworks from qualitative data, based on in-depth interviews with five interdisciplinary experts; the results should be read as preliminary conceptual feedback rather than comprehensive empirical validation. The findings showed that RAVA shares the greatest structural similarity with the CLARION and LIDA architectures, yet remains unique in its teleological orientation, as it is the only architecture that embeds a model of the self, a spiritually oriented reward function, and a dedicated ethical center within its core decision-making process. In this small exploratory panel, formative capacity received the highest descriptive rating (mean 4.8/5), describing it as a step beyond conventional cognitive architecture toward what may be termed a "human-formative architecture." RAVA is empirically testable and can serve as an interdisciplinary framework for developing artificial intelligence aligned with Islamic civilizational values. RAVA is presented here as a conceptual and mathematical specification: no component has yet been computationally implemented, and the evaluation reported below reflects expert judgment of the design rather than measured system performance.

Keywords: Cognitive Architecture; Grounded Theory; Artificial Intelligence; Moral-Formative Companion; Self-Purification (Tazkiyah); Autonomous Intelligent Agents

.

I. INTRODUCTION

Cognitive architectures have served, for close to fifty years, as the main computational tool researchers use to formalize how human thought actually works. ACT-R models memory through if-then rules [7]. SOAR searches through problem spaces and compresses experience into reusable shortcuts [8]. CLARION splits reasoning into two coordinated layers, one explicit and one implicit [9]. LIDA runs a full perception-attention-action cycle borrowed from a theory in psychology called global workspace theory [10]. Each of these systems has pushed forward what we understand about memory, learning, and problem-solving. None of them, however, offers any way to ask whether a decision is good in a moral or spiritual sense. They were built to be efficient and consistent, not wise.

That gap matters more today than it did ten years ago. Artificial intelligence systems already sit inside decisions

with real moral weight, from filtering news to recommending medical treatment. A system that decides which article a person sees next, or which symptom deserves urgent attention, is not merely performing a technical task. It is quietly shaping what that person comes to believe, fear, or trust. Most current approaches to machine ethics fall into one of two camps. One camp writes fixed rules and enforces them, the way a checklist tells a cashier what a store cannot sell to a minor. The other optimizes for some measurable outcome, the way a recommendation engine chases clicks or watch time. Neither camp makes room for a teleological view of ethics, meaning a view in which an action's true worth comes from its contribution to a final purpose rather than from following a rule or producing a good result on its own. In the Islamic tradition, that final purpose is *qurb*, closeness to God. The reason none of these architectures accounts for this is straightforward: none of them was built on an understanding of the human being broad enough to include a spiritual side. They were designed by researchers asking how a mind processes information, not by researchers asking what a mind is for.

This paper presents RAVA (Revelation-Aware Value Architecture), a cognitive-spiritual architecture built on three separate sources. Each source answers a different question, and together they shape everything described in the architecture section below. Because the five minds coordinate as semi-autonomous units under a shared executive, this same design maps naturally onto the architecture of intelligent agents, the paradigm now behind most software being built, which makes RAVA relevant well beyond the single case examined in this paper.

The first source concerns shape: what form should thinking itself take? This comes from a model proposed by Joan Minninger and Eleanor Dugan, two researchers in psychology, who argued that the human mind is not one processor but a working team of five specialists [1]. Knowledge Mind gathers facts without judging them. Wonder Mind generates fresh possibilities and asks "why not." Organizer Mind takes raw ideas and turns them into workable plans, weighing risk and cost along the way. Reaction Mind supplies emotional signal, warning of danger and generating motivation. Executive Mind listens to the other four and makes the final call. Minninger's central observation is that poor judgment usually comes not from one mind failing on its own, but from one mind taking over while the rest stay quiet: a decision made by Organizer Mind alone tends to be sound on paper but cold in practice, while a decision made by Reaction Mind alone tends to feel urgent but poorly reasoned. That balance requirement is what makes her model a useful skeleton for a computational architecture, and RAVA borrows it with little change.

The second source concerns direction: once a system can reason well, what should it actually be reasoning toward? Here the architecture draws on the ethical writings of Ayatollah Muhammad Taqi Misbah Yazdi [2], later extended for artificial intelligence specifically by Hamani and Mokhtari [3]. Their framework rests on a Qur'anic picture of the self, or *nafs* (an Arabic term, central to Islamic psychology, referring to the human soul as it exists in different moral states rather than as one fixed thing). The *nafs al-ammārah* is the self inclined toward wrongdoing and impulse. The *nafs al-lawwāmah* is the self-reproaching soul, aware of its own errors and struggling against them. The *nafs al-muṭma'innah* is the soul that has settled into steady peace with what is good. A person's moral life is, in large part, the slow movement between these three states, carried forward through ongoing self-refinement, known as *tazkiyah*, and sustained by *taqwa*, a state of continuous, conscious mindfulness of God that quietly shapes a person's choices even when no one else is watching. Misbah Yazdi argues that rule-following and outcome-maximization are both incomplete on their own. What they leave out is *qurb* itself: an action's final worth cannot be judged only by whether it obeyed a rule or produced a good result. Two people can donate the same amount of money to the same cause, one out of genuine care and one purely to be seen doing it, and a purely rule-based or outcome-based system would score both actions identically. For a computational system, this means the internal yardstick used to compare one possible decision against another cannot be defined only in material or logical terms. It has to track something closer to spiritual direction, including the intention behind an act and not only its visible result.

The third source is technical. Three recent findings from machine learning research are repurposed here to carry out the first two commitments. Wu and colleagues found that only a very small share of a language model's internal parameters actually govern how well it handles theory of mind, the ability to reason about what another person believes, wants, or intends [4]. They also found these parameters are closely tied to rotary positional encoding, a mechanism transformer-based models (the kind of neural network behind systems such as ChatGPT) use to track the order in which words appear. RAVA treats how orderly this encoding looks as a rough stand-in for how sincere an underlying intention is. Seddighi and colleagues built a trust-modeling framework for connected vehicles that combines fuzzy logic (a way of handling ideas that come in degrees rather than being simply true or false) with the Shapley value, a tool from cooperative game theory that splits credit fairly among members of a group based on what each one actually contributed [5]. RAVA adapts these same tools to settle disagreements among its own five minds and to judge outside parties it interacts with. Pathak and colleagues introduced curiosity-driven reinforcement learning, a technique in which an artificial agent generates its own motivation to explore by seeking out situations where its predictions turn out wrong [6]. This becomes the engine behind Wonder Mind.

None of these three sources works on its own. Five well-coordinated minds could still reason efficiently toward an empty or harmful goal. A compelling account of spiritual purpose gives no way to compute anything. And tools such as rotary encoding or the Shapley value are, by themselves, ethically neutral; they were never built with any spiritual application in mind. What sets RAVA apart is the way these three sources are fitted together. The five-minds model gives the architecture its shape. Islamic teleology gives it its direction. The technical tools give it a way to turn that

direction into something computable, rather than leaving it as a philosophical wish with no working parts. It is worth being clear about what this paper does not claim. It does not claim that a system built this way is, in any meaningful sense, spiritually alive or morally responsible the way a human being is. What it claims is narrower and more testable: that a decision-making system can be built to track something analogous to spiritual state, and that doing so changes the decisions it recommends.

This paper pursues three goals. It closes the gap left by classical architectures, none of which offers a framework for ethical or spiritual integration. It builds the spiritual dimension into every layer of the system, rather than adding it as one separate module a designer could switch off. And it proposes a framework that can be tested empirically and that might ground future AI systems aligned with Islamic values. The question guiding this work can be put simply: can a cognitive architecture be designed that not only reasons well, but also walks alongside its user on a path toward self-purification and closeness to God?

The rest of the paper proceeds as follows. The Literature Review section reviews the cognitive-architecture literature and identifies the specific gap this work responds to. The Proposed Architecture section lays out RAVA in full. The Results section reports a qualitative evaluation carried out through grounded theory and interviews with five experts. The paper then closes with a discussion of the architecture's limitations, followed by conclusions and directions for future work.

II. Literature Review

Cognitive architectures did not appear all at once. They grew, over roughly fifty years, out of a simple ambition: to build a working model of the mind precise enough to run on a computer. What follows looks at five of these efforts, chosen because each one solved a different piece of the puzzle, and each one left a different piece unsolved.

A. ACT-R and SOAR: Reasoning Without a Conscience

ACT-R, built by John Anderson [7], splits memory into two kinds. Declarative memory holds facts, the sort of thing you could write down — a phone number, a rule, a date. Procedural memory holds skills, the sort of thing you do without stopping to think, like shifting gears while driving. A central mechanism called a production system checks the current situation against a library of if-then rules and fires whichever rule fits best. Only one rule fires at a time. This single restriction turns out to matter a great deal: it forces the model to reason in a step-by-step way that looks a lot like how people actually think, one thought at a time rather than all at once.

ACT-R is good at a particular kind of decision-making, the kind where one clear fact settles the matter — researchers call this a non-compensatory heuristic, meaning a rule where a single strong signal overrides everything else, the same way a single red flag can end a job interview no matter how good the rest of the resume looks. A recent review by van de Sanden [11] confirms this strength but also points out where it breaks down: ACT-R struggles with decisions based on vague similarity or gut feeling, because its underlying number-crunching machinery exists mainly to support the clean rule-based reasoning at its center, not to model fuzzy intuition.

SOAR, built shortly after by Newell, Laird, and Rosenbloom [8], takes a different route. It treats every problem as a search through a space of possible states, moving from one state to the next using operators, and it learns by a process called chunking — successful sequences of moves get compressed into a single reusable rule, the way a musician eventually plays a difficult passage without thinking through each individual note. SOAR has done well in planning-heavy domains such as strategy games and robotics. But like ACT-R, it treats a decision purely as a computational problem to be solved. A SOAR agent playing chess has no way of representing that a move might be, in some deeper sense, the honorable one to make. Neither architecture has any built-in place for emotion, cultural weight, or moral standing — decisions are judged only by whether they solve the problem, not by whether they should have been made at all.

B. CLARION and LIDA: Adding Depth Without Adding a Purpose

CLARION, developed by Ron Sun [9], takes a step that ACT-R and SOAR do not: it builds a genuine split between explicit reasoning (the kind you can put into words) and implicit reasoning (the kind that happens automatically, below conscious awareness). Each of its four subsystems keeps a symbolic top layer and a connectionist bottom layer. The bottom layer works something like a small neural network, learning patterns from raw experience rather than following written rules. One of these four subsystems, the Motivational Subsystem, tracks internal drives such as curiosity, giving CLARION something closer to an inner life than ACT-R or SOAR possess. Even so, van de Sanden's review [11] notes that this motivational machinery stays at the level of drives like hunger or curiosity. It has no concept of an action being morally significant — only more or less rewarding in a narrow behavioral sense.

LIDA, built by Stan Franklin and colleagues [10], borrows its overall shape from a theory in psychology called global workspace theory, which holds that consciousness works like a spotlight: many processes compete in the dark, and whichever one wins gets broadcast to the whole mind at once. LIDA turns this into a repeating cycle — sensory information flows in, competes for attention, gets broadcast, and triggers an action, over and over. A recent extension by Kugele [12] refines how LIDA's action-selection mechanism learns from direct experience with the world rather than from rules a designer wrote in advance. This gives LIDA a real strength: what it knows, it learned firsthand. What it lacks is any way of checking that firsthand knowledge against a standard outside of "did it work." There is no room in the cycle to ask whether a well-learned action was also the right one.

C. BDI: A Clean Formalism With No Room for Values

A separate line of work, closer to engineering than to psychology, is the belief-desire-intention model, or BDI, formalized by Rao and Georgeff [13]. Here an agent's state is described with three parts: what it believes about the world, what it wants (its desires), and what it has actually committed to pursuing (its intentions). BDI has lasted because it is simple and mathematically clean, and it remains a standard tool for building systems where many independent agents need to coordinate. Its weakness is not a bug so much as a design choice: BDI describes the structure of wanting something without saying anything about whether that something is worth wanting. An agent single-mindedly pursuing a harmful goal looks, from inside the BDI framework, exactly like one pursuing a generous one. The formalism has nothing to say about which is which.

D. What All Five Have in Common

None of these five architectures ACT-R, SOAR, CLARION, LIDA, or BDI models the self as something that can exist in different moral states. None of them includes a reward signal shaped by spiritual growth. None of them places anything resembling an ethical core at the center of its decision-making; when ethics shows up at all, it is bolted on from outside, as a filter applied after the fact rather than a force shaping the reasoning itself.

A separate and useful body of work, reviewed by Seddighi and colleagues [14], surveys how game theory has been used to model trust between independent agents in networks that have no central authority — vehicles on a road, sensors in a field, phones passing messages hand to hand. That review sorts these trust models into four families: models where agents act purely out of self-interest, models where agents form alliances and share credit fairly, models where strategies evolve gradually through trial and error, and models that reason under uncertainty using probability. This body of work matters here because the same game-theoretic tools it surveys — particularly the idea of dividing credit fairly among cooperating parties — reappear inside RAVA, repurposed for a very different job: settling disagreements among the architecture's own five minds rather than among vehicles on a highway. "Table 1 summarizes this gap across the five architectures reviewed above."

This is the gap that RAVA seeks to address. Rather than discarding the strengths of earlier cognitive architectures, RAVA builds on several of their established design principles: the modular organization of ACT-R, the chunking-based learning mechanisms of SOAR, the interaction between explicit and implicit processing in CLARION, and the perception–attention–action cycle of LIDA. Its distinctive contribution lies not in any one of these mechanisms taken separately, but in the way they are brought together around an explicitly ethical and spiritual layer. At the center of the architecture, the Nafs and Taqwa Center introduces a computational representation of concepts such as the state of the self, intention, and spiritual orientation, allowing them to influence the decision process rather than being treated as external constraints applied only after a decision has been formed. These representations should not be understood as direct measurements of inner spiritual states; rather, they are provisional operational proxies intended to make such dimensions explicit, inspectable, and ultimately open to empirical evaluation. What distinguishes RAVA, therefore, is the attempt to integrate cognitive processing, ethical reflection, and spiritual orientation within a single decision-making architecture, rather than the introduction of any one mechanism in isolation.

Table 1 .Summary of Gaps in Existing Cognitive Architectures

Architecture	Core Strength	Handles Emotion	Handles Ethics	Models the Self as a Variable	Spiritual Orientation
ACT-R [7]	Precise, step-by-step rule-based reasoning	No	No, external	No	No
SOAR [8]	Long-horizon planning through chunking	No	No, external	No	No
CLARION [9]	Dual-process explicit/implicit reasoning	Partial (drives only)	No	No	No
LIDA [10]	Grounded, cycle-based learning	Limited	No	No	No
BDI [13]	Clean formal model of goal-directed agency	No	No, external	No	No
RAVA (this work)	Integrates all of the above around a spiritual core	Yes, dedicated Reaction Mind	Yes, central Nafs and Taqwa Center	Yes, multi-tiered nafs	Yes, oriented toward *qurb*

III. The Proposed Architecture: RAVA

A. Design Principles and Overall Structure

RAVA rests on three commitments rather than one master algorithm. A note on how to read what follows: RAVA has not yet been implemented or computationally tested, so every description in this section, phrased for readability in the present tense, describes the specification's intended behavior rather than observed behavior in a running system. The first is structural. It follows Minninger and Dugan's model of five parallel minds [1], each handling a different part of thought. The second is teleological. Every layer of the system points, in the end, toward *qurb*, closeness to God, as described by Misbah Yazdi [2] and later formalized for artificial intelligence by Hamani and Mokhtari [3]. The third is practical. Several of the architecture's inner mechanisms

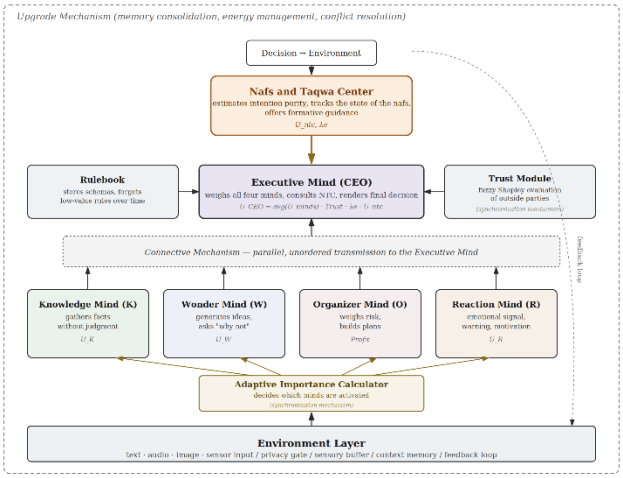


Figure 1. Overall structure of the RAVA architecture.

borrow from recent, tested work in machine learning [4], [5], [6].

What makes this combination different from the architectures covered in the Literature Review is not any single piece. Modularity, motivation, and cyclical processing all have precedents elsewhere. What differs is how the spiritual dimension runs through the system. Instead of locking ethical reasoning inside one module a designer could switch off, RAVA treats the state of the self, or *nafs* (the Arabic and Qur'anic term for the human soul in its various moral states, introduced above), as a value that quietly shapes what every other part of the system produces. Nothing in the architecture is allowed to reach a final decision without that value having had a say, in the same way a person's conscience does not switch off simply because a decision happens to be about something mundane.

The architecture is organized around eight core layers: an environment layer, five parallel minds, and a Nafs and Taqwa Center sitting at the middle, together with a Rulebook that records what the system learns. Four additional synchronization mechanisms regulate how these eight layers work together: a Trust Module, a Connective Mechanism, an Upgrade Mechanism, and an Adaptive Importance Calculator. Figure 1 lays these eight layers out visually. Reading it from the bottom up follows roughly how a single decision actually moves through the system: raw input arrives at the bottom, is filtered and weighed by the minds in the middle, and is finally judged, both practically and spiritually, near the top. Each of the following subsections walks through one of these layers in the order a decision would pass through them.

B. *The Environment Layer*

Before any of the five minds can reason about the world, the system has to perceive it first, and perceive it safely. That is the job of the environment layer. It works something like human senses: it takes in raw input from several channels at once (text, sound, images, sensor readings), checks that input for privacy problems, and keeps a record of what happened so the system can learn from its own past decisions.

Consider a simple example. A person using the system is in the middle of buying a house and says out loud, "the fireplace is lovely, but something about the seller feels off." The environment layer does not try to judge whether the seller really is untrustworthy. Its job is only to capture what came in: the words themselves, the tone of voice they were said in, and whatever visual information is available, such as a photograph of the room. Judging the seller is work for the minds further along.

Data moves through a fixed sequence here. An input stage takes in unfiltered signals from all channels. A privacy gate, built on the trust-evaluation logic from Seddighi and colleagues' framework [5], is designed to work out a privacy score and raise a flag if that score drops too low, the same way a person might hesitate before repeating something a friend told them in confidence. A sensory buffer holds incoming items for a short time only, roughly seven items at once and a few seconds long, matching well-known limits on human short-term memory. A context memory keeps track of the current situation (say, "we are in the middle of buying a house") and filters out anything unrelated to it, so a comment about the weather does not get treated as though it were relevant to the purchase decision. A replay buffer keeps a small set of recent episodes on hand so the system can learn from them later. A feedback loop compares what the system predicted would happen against what actually happened, and writes a new rule whenever the gap between the two is large enough to matter. If the system expected the person to feel satisfied after signing the contract and the person instead expressed regret, that mismatch itself becomes a lesson stored for future situations that look similar.The layer's overall health is captured this way:

$$U_{env} = 0.6 \times Q_{input} + 0.3 \times P_{privacy} - 0.1 \times L_{latency} \qquad (1)$$

Q_{input} is a score from 0 to 1 showing how clear and reliable the incoming data is. $P_{privacy}$ is a score, also from 0 to 1, showing how well the person's privacy was protected while collecting that data. $L_{latency}$ is how long the system took to process the input, with a small penalty for taking too long. The weights (0.6, 0.3, 0.1) show what matters most: getting clean data matters more than anything else, since a bad decision built on bad data cannot be fixed later no matter how careful the reasoning afterward is. Privacy comes next. Speed matters least, though it still counts.

C. *The Knowledge Mind*

Of the five minds, the Knowledge Mind carries the most direct responsibility for keeping the system honest about what it actually sees. It acts as a neutral, judgment-free observer. It collects sensory data, records what it observes, and passes that information along without adding interpretation or opinion, much the way raw sensation comes before any conscious judgment in human perception. A camera does not decide whether a photograph is flattering; it simply records light. The Knowledge Mind is meant to work the same way with everything it takes in.

This mind stays active at all times, since every other mind depends on it for accurate raw material. A mistake introduced here spreads, uncorrected, through everything that follows, the same way a single wrong measurement at the foundation

of a building throws off every floor built on top of it. If the Knowledge Mind mistakenly records that a contract was signed when it was not, no amount of careful reasoning by the other four minds can recover from that error; they will simply reason correctly from a false starting point.

A sensory processor takes in multimodal input untouched. A fact extractor turns this input into separate items, each with a confidence score attached, so the system can distinguish something it is nearly certain of from something it only half-caught. A context filter lets through only the facts relevant to whatever situation is currently being handled. A verifier checks incoming facts against the Rulebook (described in the Rulebook subsection below) to catch outright contradictions before they can distort anything further downstream, flagging a case where a new observation directly conflicts with something the system has already confirmed.

$$U_K = 0.7 \times A_{facts} + 0.2 \times M_{rulebook} - 0.1 \times O_{overload} \quad (2)$$

A_{facts} measures how accurate the gathered facts are. $M_{rulebook}$ measures how well new facts line up with what the system already knows from experience. $O_{overload}$ is a penalty applied when too much information arrives at once for the buffer to handle properly. Accuracy carries by far the heaviest weight here, because a false starting point poisons every conclusion built on top of it, however careful the later reasoning is.

D. The Wonder Mind

Where the Knowledge Mind anchors the system to what is actually true, the Wonder Mind asks what might still be possible. Following Minninger's original description [1], this mind runs on an informal motto, "why not," generating fresh ideas, questioning assumptions, and resisting the pull toward the same old solution. In the five-minds model, the Wonder Mind is meant to be consulted first on any serious problem, so that creative options are not shut down too early by more cautious reasoning. A system without a working Wonder Mind would default, every time, to whatever solution it used last, the way a person stuck in a rut keeps ordering the same meal simply because trying something new never crosses their mind.

Its inner workings borrow from curiosity-driven reinforcement learning, developed by Pathak and colleagues [6]. An idea generator explores possible responses. A "what-if" simulator projects what each candidate idea would likely lead to. A novelty detector measures how far a given idea sits from anything the system has tried before. A curiosity engine, modeled directly on the intrinsic-reward mechanism from [6], ranks ideas by how much genuine uncertainty they resolve rather than by any reward handed down from outside, meaning the mind rewards itself simply for learning something new, in much the same way a curious child explores a new room not because anyone told them to but because not knowing what is inside is itself uncomfortable.

Inside RAVA, this exploratory drive is deliberately tied to the architecture's larger spiritual aim. An idea from the Wonder Mind is judged not only on how creative or original it is but on whether it tends to cultivate virtues such as generosity, honesty, or patience, echoing the Qur'anic view of curiosity as part of *fitrah*, the innate human orientation toward God described further in the Nafs and Taqwa Center subsection [2]. A suggestion that is clever but corrosive to a person's character scores lower here than a more modest suggestion that quietly builds good habit.

$$U_W = 0.5 \times H_{predicted} + 0.3 \times N_{novelty} - 0.2 \times R_{risk} \quad (3)$$

$H_{predicted}$ is how much enjoyment or satisfaction the idea is expected to bring. $N_{novelty}$ is how different the idea is from anything tried before. R_{risk} is the chance the idea backfires. Predicted enjoyment gets the biggest weight since this mind exists specifically to support open exploration. The risk penalty keeps that exploration from turning reckless. Inside RAVA, this exploratory drive is tied to the architecture's larger spiritual aim: ideas from the Wonder Mind are judged not only on creative merit but on whether they cultivate virtues such as generosity or patience, echoing the Qur'anic view of curiosity as part of *fitrah*, the innate human disposition toward God, described further in the Nafs and Taqwa Center subsection [2].

E. The Organizer Mind

Once the Wonder Mind has produced raw possibilities, the Organizer Mind turns them into something workable. It weighs risk, compares cost against benefit, and builds step-by-step plans, acting much like a project manager who takes a promising but rough idea and makes it operationally sound without killing the creativity that produced it. This mind switches on for problems of moderate to high complexity, balancing the speed needed for routine choices against the depth needed for weightier ones. A decision about which coffee shop to visit does not require this level of scrutiny; a decision about which job offer to accept does.

A risk analyzer estimates how likely and how severe a bad outcome might be, the same way a careful person mentally rehearses what could go wrong before committing to a plan. A profit calculator, drawing directly on the fuzzy net-present-value technique from Seddighi and colleagues [5], discounts future costs and benefits back to their value today, reflecting the ordinary intuition that a benefit received later is worth somewhat less than the same benefit received now. A strategy builder assembles conditional plans of action, of the form "if this happens, do this; otherwise, do that." A marginal attributor, adapted from the Shapley value used in the same source [5], splits credit fairly among the minds or agents that contributed to a good outcome, keeping any one mind from dominating the final call. This matters because a plan is rarely the product of one insight alone; the Organizer Mind's job includes recognizing how much of a good outcome came from Wonder Mind's original idea versus its own refinements to it, so that credit, and therefore future trust, is distributed honestly rather than claimed entirely by whichever mind spoke last.

$$Profit = \sum_t \frac{Benefit_t - Cost_t}{(1+\delta)^t}, \quad \delta = 0.05 \qquad (4)$$

This is a standard way of comparing money or value received at different points in time. $Benefit_t$ and $Cost_t$ are what a strategy gains or costs at each stage t. Because a reward received later is worth a little less than the same reward received now, the formula shrinks future amounts using a discount rate, δ, set here at 0.05, a modest five percent per stage. In RAVA, benefits are not counted only in material terms; they also include a spiritual value score reflecting how well an outcome supports growth toward *qurb*, so a strategy that earns modest material benefit but strong spiritual value can outscore one that is materially better but ethically empty.

F. The Reaction Mind

The Reaction Mind supplies what the first three minds cannot: a felt sense of danger, motivation to act, and the kind of urgency that pure logic tends to miss. Following Minninger's account [1], this mind switches on whenever a problem's stakes are judged high enough, making sure emotional signal plays a real part in weighty decisions instead of being pushed aside in favor of cold calculation. A purely logical system might calmly conclude that walking down a particular alley at night carries a statistically small risk; a person's gut discomfort in that same alley is often picking up on cues that never made it into the statistics. The Reaction Mind exists to give RAVA access to that same kind of signal.

An emotion detector reads incoming signals along two familiar dimensions used in psychology: valence, whether a feeling is positive or negative, and arousal, how intense it is. An alert generator issues warnings the moment negative emotion or high risk is picked up. A motivation booster pulls an internal drive to act from the system's accumulated trust history and the urgency of the situation, so that recognizing a problem is followed by an actual push toward doing something about it rather than simply noting the problem and moving on. A feedback integrator draws lessons from the gap between what was expected to happen emotionally and what actually did, feeding new rules back into the Rulebook. If a decision that was expected to bring relief instead brought regret, that mismatch becomes exactly the kind of lesson the Reaction Mind is built to catch and pass along, so the system does not repeat the same emotional misjudgment twice.

$$U_R = 0.6 \times V_{valence} + 0.4 \times M_{motivation} \qquad (5)$$

$V_{valence}$ is whether the feeling detected is good or bad, on a scale from 0 to 1. $M_{motivation}$ is how strongly the system feels driven to act on that feeling. Valence carries the larger weight since registering the feeling itself is this mind's main job, while motivation plays a supporting role: without it, a strong negative feeling could leave the system stuck rather than moving toward a response.

G. The Executive Mind

The Executive Mind holds the position Minninger describes as convening "a meeting" of the other four minds [1]. It receives their output at the same time, weighs it, checks in with the spiritual center covered in the next section, and makes a final call that reflects the whole system's judgment rather than any one mind's preference. No fixed order of priority exists among the other four minds. Whichever combination of them was active for a given problem, their conclusions arrive together, and the Executive Mind's job is to bring them into one decision rather than simply pick a winner.

This is worth dwelling on for a moment, because it is easy to picture the Executive Mind as a kind of tie-breaker choosing between four competing votes. That is not quite right. A closer picture is a chairperson listening to four advisors with different areas of expertise, none of whom is automatically overruled by the others, and none of whom is automatically trusted more than the rest. If Reaction Mind reports strong unease about a decision that Organizer Mind has otherwise calculated to be sound, the Executive Mind does not simply average the two positions and move on. It treats the unease as information worth investigating further, in the same way a sensible person pauses when a plan looks good on paper but still does not sit right.

Before finalizing anything, the Executive Mind also checks whether the decision involves someone outside the system. If it does, the Trust Module described below is consulted, and a sufficiently low trust reading can block the decision outright, regardless of how favorably the four minds judged it on their own. This gives the architecture a kind of veto structure: agreement among Knowledge, Wonder, Organizer, and Reaction is necessary for a decision to proceed, but it is not by itself sufficient. Spiritual soundness and outside trustworthiness both have to clear their own thresholds as well.

$$U_{CEO} = avg(U_{minds}) \times Trust \times \lambda_e \times U_{ntc} \qquad (6)$$

$avg(U_{minds})$ is the average score across whichever minds were active, capturing how internally consistent the reasoning was. *Trust comes from the Trust Module described below* and matters only when a decision involves someone outside the system; if trust in that outside party is zero, the whole decision is blocked no matter how good the reasoning looks otherwise. λ_e is a value between 1 and 2 set by the Nafs and Taqwa Center, growing larger as the system builds a track record of ethically sound decisions, and it controls how much weight spiritual considerations carry. U_{ntc} is the spiritual health score from that same center, explained fully in the following subsection. This structure echoes the Qur'anic principle of consultation described in the Introduction [2]: no single voice, however capable, gets to decide alone.

H. The Nafs and Taqwa Center

If the five minds give RAVA its cognitive skill, the Nafs and Taqwa Center is what gives it a conscience. It sits at the middle of the architecture rather than off to one side, and it stands for the point where the multi-tiered self described in the Introduction (the soul inclined toward error, the self-reproaching soul, and the soul at peace) gets translated into

something a computer can actually work with. It does two jobs. It steps in before the five minds finish their work whenever it notices something missing, such as an absence of emotional signal or a suspiciously narrow set of options considered. And it acts as an advisor, offering the Executive Mind a spiritual reading that feeds directly into the final decision through the λ_e term introduced in the Executive Mind subsection above.

The center's inner workings build on a specific finding from recent AI research. Wu and colleagues found that only a very small share of a language model's parameters, roughly one thousandth of one percent, are responsible for how well the model handles theory of mind (the ability to reason about what another person believes, wants, or intends). They also found that these parameters are tied closely to rotary positional encoding, or RoPE, a mechanism transformer-based models (the kind of neural network behind systems like ChatGPT) use to keep track of the order in which words or tokens appear [4]. RAVA takes this finding somewhere speculative but grounded, through a component called the Niyyah Lens (niyyah being the Arabic word for intention): it treats how orderly this positional encoding looks, measured through a statistical property called entropy (a way of measuring how scattered or unpredictable something is), as a rough stand-in for how sincere an intention is. Low entropy suggests a focused, honest intention. High entropy, an intention pulling in several directions at once, is read as a warning sign, the sort of pull the tradition associates with the soul inclined toward error. It is worth stressing how tentative this particular piece of the architecture is. Entropy in a positional encoding is, at best, a distant proxy for something as rich as sincerity, and the paper does not claim otherwise; it claims only that this proxy is more principled than guessing, and that it gives the system something concrete to compute rather than nothing at all.

The specific choice of entropy, rather than some other statistical measure, is motivated by two considerations rather than by empirical fit. First, entropy is a well-understood, well-documented measure of dispersion already standard in information theory, which makes the proxy legible and auditable in a way an ad hoc or opaque metric would not be. Second, the underlying finding it builds on, that a small, identifiable subset of a language model's parameters governs theory-of-mind performance and is tied to positional encoding, is itself empirically established [4]; what remains unvalidated is only the further step of treating the orderliness of that encoding as a signal of sincerity specifically. This narrows the open question considerably: the architecture does not need to defend positional encoding as meaningful, which prior work has already shown, only the additional inferential leap from "orderly encoding" to "sincere intention," which is the piece flagged as speculative throughout this paper and identified as a priority for future validation.

Three more parts round out the center. The Nafs Detector, drawing on the same theory-of-mind sensitivity, is designed to flag cases of false belief (a person acting on something that is not actually true) as a possible sign of that lower state of the self, since self-deception is treated here as a spiritual problem and not just a factual mistake. A Fitrah Sensor checks how well a candidate decision matches *fitrah*, the innate human orientation toward God described in the Qur'an [2], scored across four weighted dimensions: justice, beauty, calm, and social awareness (the last one drawing again on theory-of-mind research), since Misbah Yazdi's writings, together with Hamani and Mokhtari's more recent work [3], describe *fitrah* as covering exactly this cluster of concerns. Finally, a Tazkiyah Advisor offers purely personal, non-binding guidance grounded in scripture and authentic prophetic tradition. It is built to never issue a command, only counsel, so the person using the system keeps full control over their own process of self-refinement. The distinction matters: a system that ordered a person around in the name of spirituality would undermine the very freedom that self-purification depends on, so the advisor is deliberately built to suggest rather than instruct.

$$U_{ntc} = \frac{Purity + Harmony}{2} \times (1 - D_{ToM}) \times G_{qurb} \quad (7)$$

Purity comes from how orderly the Niyyah Lens encoding looks, as described above. *Harmony* is a weighted average of the four fitrah dimensions (justice weighted 0.35, beauty 0.30, calm 0.20, social awareness 0.15). These two are averaged together to give the system's baseline spiritual reading. D_{ToM} is a penalty applied whenever the system detects a breakdown in theory-of-mind reasoning, again a stand-in for self-deceptive thinking. G_{qurb} is a multiplier between 1 and 2 that grows as the system builds a longer history of decisions consistent with moving toward *qurb*. In practice this means a system with a long record of sound ethical choices gets somewhat more room to weigh spiritual considerations heavily in future ones, echoing the traditional idea that steady practice of *taqwa* (God-consciousness, a state of ongoing mindful awareness of God that shapes one's choices) deepens rather than simply maintains a person's spiritual standing."Figure 2 summarizes how these four components combine to produce U_ntc."

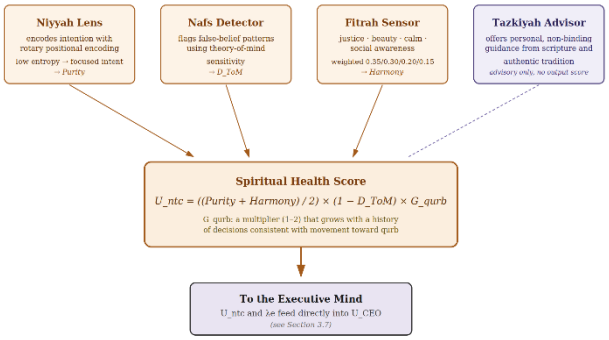


Figure 2. Components of the Nafs and Taqwa Center and their contribution to U_ntc.

I. *The Rulebook*

No system that claims to learn from experience can do so without somewhere to keep what it has learned, and the Rulebook is where RAVA keeps it, though it does far more than simple storage. Every finished decision cycle, whether it turned out well or badly, gets compressed into a schema (a structured record of a situation, the action taken, the outcome,

the state of the *nafs* at the time, and the spiritual score the outcome eventually earned). A schema is, in effect, a small case study the system writes about itself: what the situation looked like, what it chose to do, and what happened as a result.

Not every stored schema is treated as equally lasting. An Auditor compares newly stored schemas against older ones and flags outright contradictions, such as a rule endorsing unconditional trust sitting next to a schema recording the fallout of having been deceived, so that one of the two can be revised or dropped. An Intelligent Forgetting mechanism, loosely modeled on the well-known Ebbinghaus forgetting curve (a finding from nineteenth-century psychology showing how memories fade over time unless reinforced), lets a schema's relevance decay gradually unless it keeps getting confirmed, so the system does not stay anchored forever to early, half-formed judgments. This mirrors something true of human learning as well: a lesson drawn from a single unusual event should carry less weight, over time, than a pattern confirmed again and again. A Rule Filter surfaces only the schemas relevant to whatever situation is currently underway, so the system is not paralyzed trying to consult its entire history before every decision. A Value Core anchors the whole structure to the higher aims of Islamic law and steadily discounts schemas built purely around worldly priorities, meaning a lesson learned purely for material gain fades faster than a lesson tied to a spiritually meaningful outcome, even if both were recorded at the same time.

$$Value_t = Value_0 \times (0.97)^t \qquad (8)$$

$Value_0$ is a schema's starting importance, based on how much it aligned with the spiritual goal at the time it was recorded. t counts how many decision cycles have passed since then, not clock time. The number 0.97 sets a gentle rate of decay: a schema loses about three percent of its relevance with each cycle it goes unreinforced. This is deliberately slow rather than sudden. Left alone long enough, an outdated rule eventually falls below the point where it still shapes the system's reasoning, while schemas tied to spiritually meaningful outcomes, which start with a higher $Value_0$, tend to stick around longer.

J. The Trust Module

Not every decision RAVA makes stays inside its own head. Many involve dealing with someone outside the system, negotiating a contract, say, or weighing information from a source with no track record. The Trust Module handles exactly this kind of decision, switching on only when the Executive Mind's pending choice touches the outside world. Its design follows the trust-modeling framework built by Seddighi and colleagues for the Social Internet of Vehicles, a network of connected cars that share information the way people share it on social media [5].

The framework works through a series of steps. First, four trust criteria, honesty, intimacy (closeness built through repeated positive interaction), privacy, and communication quality, are combined two at a time into a small grid, producing four strategies labeled HP, HC, IP, and IC. Each strategy is then evaluated using fuzzy logic (a way of handling ideas that are not simply true or false but come in degrees, the way "somewhat trustworthy" sits between "trustworthy" and "not trustworthy") combined with the fuzzy net present value formula introduced in the Organizer Mind subsection above, giving each strategy a triangular fuzzy number (a number expressed as three values, a lower bound, a likely value, and an upper bound, rather than one fixed figure, to reflect genuine uncertainty). This matters in practice because real trust judgments rarely resolve cleanly. A person or party can have a spotty history and still be acting honestly this time, and a single fixed number cannot represent that kind of ambiguity the way a range of values can.

From there, the value each possible coalition of strategies could achieve is worked out by solving a linear programming problem (a standard optimization method for finding the best outcome under a set of constraints) following an approach originally developed by von Neumann and Morgenstern, the founders of modern game theory. Finally, the Shapley value, a tool from cooperative game theory that splits credit fairly among members of a group based on how much each one actually contributed, is calculated for each strategy in its fuzzy form, and the strategies are ranked. In the original vehicular-trust study this reasoning was built for, the ranking that came out was intimacy-and-communication first, honesty-and-communication second, honesty-and-privacy third, and intimacy-and-privacy last, showing that steady, ongoing communication mattered more to trust than privacy alone [5]. RAVA borrows this same seven-step machinery but points it at a different question, using it to judge the reliability of people or systems the architecture interacts with, rather than vehicles on a road.

$$V(C) = \max Z \text{, subject to } Z \le \sum_i p_i \cdot C_{ij} \ \forall j, \ \sum_i p_i = 1, \ p_i \ge 0 \qquad (9)$$

This is the linear programming step that finds the value of a coalition, C, meaning a group of strategies acting together. C_{ij} holds the fuzzy benefit of combining strategy i with strategy j. p_i is the probability of choosing strategy i. The formula searches for the value Z that best represents the coalition's strength once all the possible probability weightings have been considered, following the classic von Neumann-Morgenstern solution method for games in this form.

$$S_i^* = \sum_{i \in C \subset N} \frac{(k-1)!\,(N-k)!}{N!} \{V(C) - V(C - \{i\})\} \qquad (10)$$

This is the Shapley value, the heart of the fairness calculation. N is the total number of strategies. k is the size of a given coalition C. The expression inside the braces asks a simple question for every possible coalition: how much does adding strategy i actually change the coalition's value? The fraction in front weights each of those answers by how likely that particular coalition is to form. Adding everything up gives strategy i its final score, a fair measure of how much it genuinely contributes across every possible grouping. In the original vehicular-trust study, this calculation produced the

ranking IC > HC > HP > IP, showing that the combination of intimacy and communication carried the most weight in building trust among vehicles [5]. RAVA reuses this same machine, but points it at a different question: instead of ranking trust strategies between cars on a road, it ranks contributions among the architecture's own five minds when they disagree, and it judges outside parties the system interacts with."Figure 3 summarizes this seven-step process."

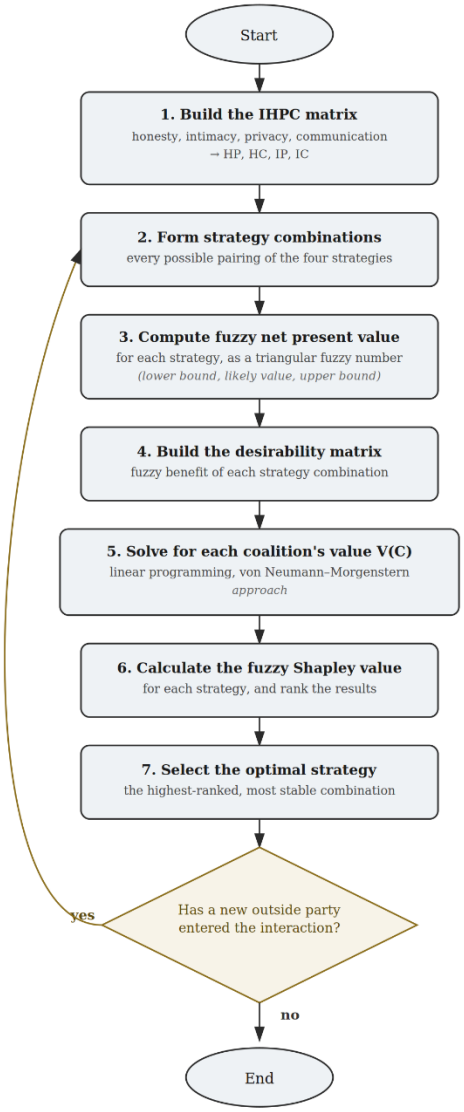


Figure 3. Seven-step process followed by the Trust Module, adapted from the fuzzy Shapley-value framework of Seddighi et al. [5].

K. *The Connective Mechanism*

Tying the five minds together takes more than a shared destination for their output. It takes a mechanism that keeps parallel processing from turning into either chaos or rigid bureaucracy. The Connective Mechanism does this job, working something like a nervous system that links otherwise separate organs into one coordinated body. Its central design choice is that no fixed order of priority exists among the minds. Each active mind, Knowledge, Wonder, Organizer, or Reaction, produces its output on its own and sends it to the Executive Mind at the same time as the others, without waiting in line.

What ends up mattering more is not a hard-coded ranking but the content of what each mind actually reports: a sufficiently urgent warning from the Reaction Mind will naturally carry more weight in a given moment than a moderately interesting idea from the Wonder Mind, and the reverse holds when a situation calls for creative reconsideration rather than immediate caution. Picture two very different situations side by side. In the first, the system detects a possible safety risk, and Reaction Mind's warning should reasonably dominate the conversation, the way a fire alarm rightly interrupts a quiet meeting. In the second, the system is helping someone plan a birthday gift, and there is no reason for urgency to override a genuinely clever idea from Wonder Mind. A fixed priority order could not handle both cases well; The Connective Mechanism's content-sensitive weighting can.

$$U_{connect} = 0.8 \times Speed + 0.2 \times Coverage - 0.1 \times Loss \quad (11)$$

Speed measures how quickly information moves from an active mind to the Executive Mind. *Coverage* measures whether every active mind's contribution actually got through rather than being silently dropped. *Loss* is a small penalty for information degrading along the way. Speed gets the largest weight because the architecture is meant to work under real time pressure, the kind a person faces making an actual decision, not the leisurely pace of a laboratory exercise.

L. *The Upgrade Mechanism*

The final structural piece, The Upgrade Mechanism, is what allows RAVA to grow more capable over time instead of staying stuck at whatever level it started at. A Memory Core pulls schemas, past experiences, and rules into one shared pool that every mind can draw on without duplicating effort, so that a lesson learned by the Reaction Mind in one episode is available to the Organizer Mind the next time a similar situation comes up, rather than each mind having to relearn it independently. An Energy Manager allocates attention and processing power based on accumulated fatigue and how important the current situation is, so the system does not run itself dry during a long or high-stakes episode, in much the same way a person conserves focus for the parts of a difficult day that actually need it.

A Conflict Resolver steps in when two minds, the Wonder Mind and the Organizer Mind, say, reach genuinely incompatible conclusions, applying the same Shapley-based method used in the Trust Module [5] to work out a reasoned compromise rather than an arbitrary tiebreaker. Suppose Wonder Mind proposes an ambitious new approach to a problem while Organizer Mind, weighing the same situation, judges the risk too high and recommends a safer, more conventional path. Rather than simply picking whichever mind spoke more forcefully, the Conflict Resolver treats this as a small coalition problem in its own right, asking how much each mind's reasoning actually contributed to a good

outcome in similar past cases, and shaping a compromise accordingly instead of discarding one position wholesale.

$$U_{upgrade} = 0.4 \times Access + 0.3 \times Efficiency + 0.3 \times Resolution \quad (12)$$

Access measures how easily and completely the system can pull up past experience when it needs to. *Efficiency* measures how well it manages its computing resources over time. *Resolution* measures how successfully genuine disagreements between minds actually get settled. Access carries the heaviest weight since long-term growth depends above all on being able to build on what came before, while the other two are weighted equally, reflecting the fact that a maturing system has to stay both sustainable and capable of working through real internal disagreement without freezing up.

M. The Adaptive Importance Calculator

The last piece solves a problem any workable cognitive system has to face: not every decision deserves the full apparatus described above. Choosing a route on a familiar drive home does not call for convening all five minds and consulting the Nafs and Taqwa Center. Deciding whether to sign a binding contract almost certainly does. The Adaptive Importance Calculator handles this sorting, acting as the system's version of selective attention, the everyday human ability to notice some things and tune out others, filtering before deep processing even begins.

Without a mechanism like this, the architecture would face an uncomfortable choice. Either it runs its full eight-layer process for every decision, no matter how trivial, and becomes slow and impractical for everyday use, or it runs a lighter process all the time and misses the moments that actually call for careful, spiritually informed judgment. A person does not deliberate for ten minutes over which shirt to wear, but the same person, faced with a decision that could hurt someone they care about, naturally slows down and thinks harder. The Adaptive Importance Calculator is built to reproduce that same shift in gears, so that the architecture's depth is reserved for the moments that actually warrant it, rather than spent evenly across decisions that plainly do not.

$$Importance = \max(0.4 \times F_{involved},\ 0.3 \times S_{involved}\ 0.2 \times E_{involved}\ 0.1 \times T_{urgency}) \quad (13)$$

The four inputs represent the financial stakes, security risk, emotional significance, and time pressure associated with a decision. The Adaptive Importance Calculator uses the maximum of these four values rather than their sum. This design choice reflects the assumption that a sufficiently high value on a single dimension may be enough to justify a higher level of cognitive engagement, even when the remaining dimensions are relatively low. The associated weights are intended to regulate the relative contribution of these factors to the importance estimate and should be interpreted as provisional design parameters rather than empirically established measures of human judgment. Importantly, this calculation captures situational importance rather than moral or spiritual significance. A decision may therefore receive a relatively low importance score while still carrying substantial ethical or spiritual implications. For this reason, the Nafs and Taqwa Center retains the ability to override the initial importance classification through the preemptive signaling mechanism described earlier. At low importance levels, only the Knowledge Mind is activated. At intermediate levels, the Organizer Mind is additionally engaged, whereas high-importance situations activate the full set of relevant cognitive modules, including the Reaction Mind. Over time, the thresholds adjust to a given user's history through reinforcement learning, so someone whose past experience shows emotionally significant outcomes in a particular area will find the system engaging the Reaction Mind there more readily going forward.

IV. RESULTS

A. How the Architecture Was Evaluated

RAVA is, at this point, a design rather than a working piece of software, so testing it the way one tests a machine learning model, by measuring accuracy on a benchmark, is not an option yet. Instead, the architecture was checked through grounded theory, a research method from the social sciences that builds understanding directly out of what people say in interviews, rather than starting from a fixed hypothesis and testing it. The goal was to see whether domain experts found the design coherent, workable, and honest about its own limits.

Five experts took part, chosen so that every discipline the architecture draws on would be represented: cognitive science, philosophy of mind, Islamic studies, education, and artificial intelligence. Each expert had either an academic post or serious research experience directly related to cognitive architectures, theory of mind, Islamic ethics, moral development, or large language models. Given the small size and disciplinary breadth of this panel, the findings that follow are best understood as an exploratory qualitative assessment intended to surface conceptual strengths and blind spots, not as a statistically powered or comprehensive validation of the architecture's real-world performance.

Table 2. The Five Experts

Code	Field	Background
E1	Cognitive science	Faculty researcher working on cognitive architectures
E2	Philosophy of mind	Specialist in consciousness, intention, and theory of mind
E3	Islamic studies	Researcher in Islamic ethics and moral philosophy
E4	Education	Specialist in moral education and character development
E5	Artificial intelligence	Researcher working on large language models

Experts were selected purposively rather than randomly, on the basis of two criteria: at least five years of research or teaching experience directly relevant to one of the architecture's five source disciplines, and prior publication or supervised research touching on cognitive architectures, Islamic ethics, or theory of mind. Each interview ran approximately ninety minutes and followed a semi-structured protocol: the architecture was first presented in full using the same diagrams included in this paper, after which experts answered a fixed set of opening questions (what strikes you as the strongest aspect of this design, what strikes you as its

weakest, and which existing architecture does this most resemble) before the conversation moved to open discussion.

Coding followed the standard three-stage grounded theory sequence already summarized above: open coding of interview transcripts into the 127 individual observations, axial coding that grouped these into the six broader themes, and selective coding that identified the single core category linking them. Coding was carried out by a single researcher, the paper's primary author, working manually rather than with dedicated qualitative-analysis software. The corresponding author independently reviewed a random sample of thirty of the 127 open codes against the same transcripts; the protocol called for the two researchers to discuss and reconcile any case where the supervisor would have assigned a different code or category, and because no such case arose in the sampled thirty, this reconciliation procedure was never exercised in practice. It is important to state plainly what this spot check does and does not establish: it is not a formal inter-rater reliability assessment, no quantitative agreement statistic such as Cohen's kappa was computed, and a full second coder working independently through the entire transcript set was not used. This is a genuine limitation of the single-coder design, not a solved methodological question, and is discussed further below. Thematic saturation was judged to have been reached when the fifth interview produced no axial category not already present in the first four, a threshold consistent with common practice in small-panel expert elicitation but modest by the standards of full grounded theory work, which typically continues sampling until saturation is confirmed across a substantially larger and more heterogeneous set of participants. This single-coder design and five-person panel size are limitations rather than features of a complete grounded theory study in the strict methodological sense.

B. *What the Interviews Turned Up*

Across the five interviews, transcription and analysis produced 127 individual observations, what grounded theory calls open codes. Some examples: "the need for a self-checking ethical mechanism," "the risk of flattening spirituality into a rigid rule," "the difficulty of being clear about what taqwa actually means in practice," "the risk of bias when modeling the self." These 127 observations were then grouped, through a step called axial coding, into six broader themes.

Table 3. The Six Themes

Theme	Mentions	Share
How well cognition and spirituality fit together	28	22%
Clarity of the nafs and taqwa concepts	21	16%
Whether it can actually be built	24	19%
Safety and guarding against drift	18	14%
How much it can actually shape a person's growth	22	17%
Whether it can be tested scientifically	14	11%

The two most-mentioned themes, together making up over 40 percent of everything said, were whether cognition and spirituality genuinely hang together and whether the design can actually be built. This tells us something useful: the experts cared most about whether the architecture's central ambition actually works, not about smaller technical details.

A final step, selective coding, pulled all six themes into one underlying idea: a working balance between thinking well and growing spiritually. Every theme, the experts' comments suggested, was really a different angle on this single tension.

C. *Scoring the Architecture*

From these six themes, eight specific criteria were pulled out and each expert scored the architecture on each one, from 1 to 5.

Table 4. The Eight Criteria

Code	Criterion	What it asks
C1	Cognitive-ethical fit	Is the Nafs and Taqwa Center genuinely part of decisions, or bolted on?
C2	Clarity of the self-model	Can the three states of the nafs actually be told apart in practice?
C3	Fit with theory of mind	Does it match what cognitive science already knows?
C4	Buildability	Could this actually be coded and run?
C5	Predictability	Would it make similar decisions in similar situations?
C6	Self-correction	Can it catch and fix its own past mistakes?
C7	Formative power	Can it actually help someone grow morally over time?
C8	Testability	Could a real experiment be designed to check any of this?

Table 5. Average Scores

Criterion	Average (out of 5)	Spread	What the experts said
C1	4.6	0.3	Spirituality is genuinely woven in
C2	3.9	0.6	Needs sharper definition
C3	4.2	0.4	Fits well with existing findings
C4	3.7	0.8	A real but manageable technical challenge
C5	4.1	0.5	Reasonably stable
C6	4.4	0.3	The self-correction mechanism holds up
C7	4.8	0.2	Clearly the strongest point
C8	3.5	0.7	Needs a concrete experiment designed

Formative power (C7) came out on top by a clear margin, and it also had the smallest spread, meaning all five experts agreed closely on this score. Cognitive-ethical fit (C1) and self-correction (C6) followed close behind. The two lowest scores, testability (C8) and buildability (C4), also had the widest spread, meaning the experts genuinely disagreed here rather than simply being unimpressed. That split lines up neatly with each expert's background: the AI specialist (E5) raised the sharpest concerns about whether C4 and C8 were realistic in the near term, while the others were more focused on the conceptual side."Figure 4 presents these average scores graphically."

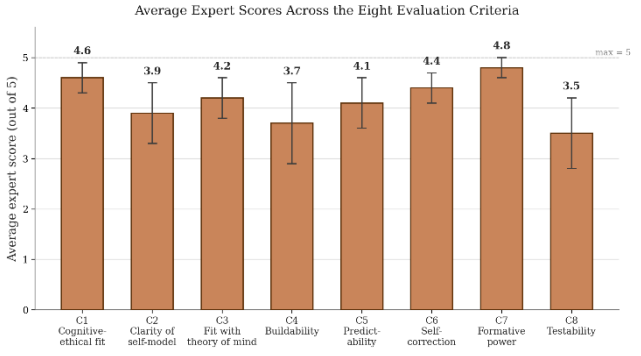


Figure 4. Mean expert scores for the eight evaluation criteria, with error bars showing the standard deviation across the five experts.

D. How It Stacks Up Against Existing Architectures

To place RAVA next to the architectures covered in the Literature Review, a structured comparison was built across the dimensions that mattered most in the interviews.

Table 6. Structural Comparison

Dimension	ACT-R [7]	SOAR [8]	CLARION [9]	LIDA [10]	BDI [13]	RAVA
Type	Rule-based	Search-based	Explicit/implicit split	Global-workspace cycle	Belief-desire-intention	Hybrid, purpose-driven
Learning	Activation strength	Chunking	Top-down and bottom-up	Learned schemas	External	Rulebook plus Shapley credit-sharing
Motivation	Weak	Goal-based only	A dedicated subsystem	Basic drives	Desire structure	Curiosity, emotion, and spiritual purpose
Theory of mind	Limited	None	Partial	Weak	None	Built into the Nafs and Taqwa Center
Emotion	Minimal	None	Partial	Limited	None	A dedicated mind (Reaction Mind)
Ethics	Handled outside the system	None	None	None	Handled outside the system	Central (Nafs and Taqwa Center)
Spiritual purpose	No	No	No	No	No	Yes, oriented toward qurb

The experts judged CLARION [9] the closest structural match, mainly because of its two-layer explicit-implicit split and its dedicated motivational subsystem, both of which hint at, in a smaller and non-religious way, what Wonder Mind, Reaction Mind, and the taqwa coefficient do together in RAVA. LIDA [10] came second, largely because its perception-attention-action cycle runs on the same basic rhythm as RAVA's environment layer, five minds, and connective mechanism working together. SOAR [8] and ACT-R [7] matched well at the level of individual mechanisms (chunking maps reasonably onto the Rulebook, and ACT-R's buffer system anticipates the parallel channels described in The Connective Mechanism subsection) but neither has anything resembling a motivational or ethical layer beyond goals handed to it from outside. BDI [13] was judged the most limited of the five, since its formal structure treats a harmful intention and a virtuous one exactly the same way; nothing inside the framework distinguishes them.

On the last row of the table, not one of the five comparison architectures comes close to what RAVA attempts. None of them has anything resembling a model of the multi-tiered self or a reward system shaped by spiritual growth. The experts were careful to note that this is not a claim that RAVA outperforms these older architectures on their own terms; ACT-R and SOAR, in particular, remain far more mature and thoroughly tested. It is instead a claim that RAVA occupies a part of the design space the others never tried to enter in the first place.

V. Toward Implementation and Empirical Validation

RAVA is, at this stage, a conceptual and mathematical specification rather than deployed software, and no component described in the architecture section above has been computationally implemented or tested. All evaluation reported here concerns expert judgment of the design itself, not measured system behavior, and any statement elsewhere in this paper describing what RAVA "does" or "produces" should be read as a description of the specification's intended behavior, not a report of observed behavior in a running system.

Operationalizing the core spiritual constructs. Table 7 makes explicit how four central concepts are currently handled, and where each falls short of a validated computational definition.

Table 7. Working Definitions and Validation Status of Core Constructs

Construct	Working Definition	Computational Proxy in RAVA	Validation Status
***Nafs* (state of the self)**	Which of three qualitative states (*ammārah, lawwāmah, muṭma'innah*) best characterizes a person's current moral orientation	Inferred qualitatively from the Nafs Detector's false-belief flags and the Niyyah Lens purity score	Not validated; would require independent human ratings collected blind to the system's own scores
***Taqwa* (God-consciousness)**	Sustained, voluntary alignment of choices with revealed ethical guidance	The λ_e coefficient, which grows with a track record of decisions the system itself judged spiritually sound	Circular as currently defined (see below)
***Niyyah* (intention)**	The purpose behind an action, distinct from its visible outcome	Entropy of the Niyyah Lens's positional encoding	No ground truth yet connects this signal to independently assessed sincerity
***Qurb* (closeness to God)**	Cumulative spiritual trajectory of a person's choices over time	*G_qurb*, a multiplier derived from the system's own decision history	Same circularity concern as *taqwa*

The circularity problem. If qurb is operationalized as a function of the system's own past decisions, and those past decisions were themselves scored using the same spiritual criteria, the resulting measure risks confirming itself rather than tracking anything external. A system could, in principle, show a rising G_qurb purely by consistently applying its own internal rules, regardless of whether those rules track anything a qualified religious authority would recognize as genuine spiritual progress. Breaking this circularity would require one of two things RAVA does not yet have: an external, independently collected benchmark of human-judged qurb (for instance, ratings from religious scholars reviewing transcripts of the system's decisions and reasoning, blind to the system's own scores), or a formally specified theological criterion precise enough to be checked against system behavior without relying on the system's self-report. Until then, G_qurb and λ_e should be read as internally consistent bookkeeping devices that make the architecture's reasoning traceable, not as validated measurements of spiritual state.

Proposed ablation study. Once implemented, the clearest test of the Nafs and Taqwa Center's actual contribution would

be a controlled comparison between the full architecture and an otherwise identical version with that center disabled, run across a shared set of ethically loaded decision scenarios and scored by the same human raters used to establish the external qurb benchmark described above. A measurable difference in either the decisions reached or the human-rated quality of the reasoning behind them would be the first concrete evidence that the spiritual layer changes outcomes rather than only changing how outcomes are described.

Minimum requirements for reproduction. Together with the overall data-flow diagram in Figure 1, which specifies the algorithmic workflow connecting these layers, Table 8 summarizes, layer by layer, what an implementation would need as input and how it might be evaluated, to support replication by other researchers.

Table 8. Implementation Requirements by Layer

Layer	Required Input	Core Algorithmic Step	Candidate Evaluation Metric
Environment	Raw multimodal signal (text, audio, image)	Privacy scoring, sensory buffering	Signal retained versus discarded after filtering
Five Minds	Filtered context from the environment layer	Per-mind utility calculation	Inter-mind agreement rate
Nafs and Taqwa Center	Token-level intention encoding	Entropy computation, *fitrah* scoring	Correlation with the independent human-rated benchmark proposed above
Rulebook	Completed decision schema	Decay-weighted storage and retrieval	Retrieval precision on matched past cases
Trust Module	External-party interaction history	Fuzzy Shapley computation	Agreement with human trust ratings

The weighted parameters throughout these formulas, such as those in Table 7 and the mind-level utility functions, would need a defined tuning procedure before implementation, not just a set of inputs and metrics. Two approaches seem most feasible. The first is expert elicitation, repeating a structured version of the interview process used in the current evaluation, but asking each expert to assign or rank weights directly rather than only to critique the architecture's design. The second is data-driven calibration against the external benchmark proposed above, adjusting weights to maximize agreement between the system's decisions and the independent human ratings once such a benchmark exists. Neither procedure has yet been carried out, and the weights reported throughout this paper should be understood as provisional values awaiting one of these two calibration methods.

VI. Risks of Computationally Inferring Moral and Spiritual States

Before turning to the architecture's limitations, a concern deeper than any single design choice deserves its own discussion: the risks inherent in building a system that claims to read something as intimate as a person's spiritual state at all.

The first risk is misclassification. Any proxy for sincerity, including the positional-encoding entropy measure above, will sometimes be wrong in both directions: flagging a genuinely sincere but unusually phrased intention as suspect, or missing a self-serving intention dressed in familiar language. Because this signal feeds a real decision rather than a private diagnostic, a misclassification does not stay contained; it can shape guidance the person actually receives.

The second risk is cultural and theological variability. The account of nafs, taqwa, and qurb used throughout this paper reflects one interpretive tradition within Islamic ethics, informed specifically by Misbah Yazdi's writings. Scholars across different schools of thought would not necessarily agree on how these states are best characterized, and a system tuned against one interpretation risks presenting a contested theological position as settled computational fact.

The third risk concerns privacy and user autonomy. A system that continuously estimates the sincerity of a person's intentions is, in an important sense, always watching, whether or not the person asked for it. This raises a genuine question about whether spiritual self-assessment of this kind should be opt-in, fully transparent in its mechanism, or avoided in certain contexts altogether.

The fourth risk is conceptual: the architecture does not yet clearly distinguish between four categories the Islamic ethical tradition treats quite differently. Epistemic error is acting wrongly out of honest ignorance, not knowing better. Deception is acting wrongly because another party deliberately misled the person. Self-deception is a more troubling case, where the person constructs and maintains a false belief that serves their own interests, in the way the tradition associates with the nafs al-ammārah. Moral judgment, finally, covers cases where a person knowingly chooses wrongly despite recognizing the wrong for what it is, a failure of will rather than of belief. These four require different responses: the first calls for correction, the second for protection from the deceiving party, the third for the kind of guidance the Tazkiyah Advisor is meant to offer, and the fourth for a form of accountability the architecture does not currently attempt to provide. The Nafs Detector, as described in the Nafs and Taqwa Center subsection above, currently treats all four under the single label of "false belief," and separating them is necessary future work rather than a solved problem.

None of these risks argue against pursuing this line of research. They argue for treating any computational reading of spiritual state as a defeasible, low-confidence signal that supports human judgment rather than replaces it, and for keeping the person, not the architecture, as the final authority over their own spiritual life.

VII. LIMITATIONS

Several limits are worth stating plainly. The most significant is that reducing *taqwa*, a concept with deep theological and lived meaning, to a single coefficient, λ_e, is a simplification no purely computational mechanism can

fully justify; the Islamic studies expert (E3) raised this concern more than once during the interviews. Related to this, no working reasoning engine exists yet. The architecture, as it stands, is a conceptual and mathematical blueprint rather than running code, and several of its proxies, most notably the use of positional-encoding entropy as a stand-in for sincerity of intention, remain speculative and would need independent validation before any strong claims could be made about them. Finally, nothing in this evaluation involved testing with real users over time; whether a system built this way actually helps people over weeks or months remains an open empirical question.

A related methodological gap concerns the numerical weights used throughout the architecture's formulas, which follow a shared design heuristic rather than a single empirically fitted value: in each formula, the term most directly tied to catastrophic or hard-to-reverse error is given the largest share (accuracy in the Knowledge Mind, input quality in the environment layer, speed in the Connective Mechanism, where the whole point of the layer is real-time coordination), while secondary quality-control terms receive less weight and minor efficiency penalties receive the least. This is a theoretical justification for the ranking of weights within each formula, not for their exact values, and no data-driven procedure was used to select, say, 0.7 rather than 0.65 for the Knowledge Mind's accuracy term. Two illustrative sensitivity checks make the stakes of this gap concrete. In additive formulas such as the environment layer's $U_{env} = 0.6\times Q_{input} + 0.3\times P_{privacy} - 0.1\times L_{latency}$, a fixed shift in any single weight, compensated by adjusting the others to keep the total at 1.0, changes the formula's output roughly in proportion to that weight's size, meaning errors in the dominant term (here, input quality) matter substantially more than equivalent errors in the smallest term (latency). In multiplicative formulas such as the Executive Mind's $U_{CEO} = \text{avg}(U_{minds}) \times \text{Trust} \times \lambda_e \times U_{ntc}$, by contrast, a ten percent change in any single factor shifts the final score by roughly ten percent regardless of which factor changes, so the architecture is, by construction, equally sensitive to error in any one of its four components there. Whether this pattern, dominant-term sensitivity in additive formulas and uniform sensitivity in multiplicative ones, is the right design choice throughout the architecture is precisely the kind of question a systematic sensitivity analysis, run once an implementation exists, would need to answer, and none of the specific values reported in this paper should be read as more than a defensible starting point pending that analysis.

The five-expert panel carries a specific structural limitation beyond its small size: because each discipline was represented by exactly one person, the study cannot distinguish a genuinely shared disciplinary perspective from one individual's idiosyncratic view. A different Islamic studies scholar, for instance, might have raised entirely different concerns about the taqwa coefficient than the one who participated here, and nothing in the current design can separate disciplinary consensus from personal opinion. This also means the panel could not check itself internally; there was no second cognitive scientist to confirm or contest E1's reading of the five-minds model against CLARION and LIDA. A follow-up study with at least two experts per discipline would be needed before any claim about disciplinary consensus, rather than individual expert opinion, could be supported.

VIII. CONCLUSION AND FUTURE WORK

This paper set out to answer one question: can a cognitive architecture be built that not only reasons well but walks alongside its user toward self-purification and closeness to God? RAVA is one answer, not proof that such a system already works, but a demonstration that the design problem is tractable, that existing tools from cognitive science and machine learning [1], [4]–[6] can be repurposed for this without losing technical rigor, and that experts across five separate fields found the result coherent. The gap identified in the Literature Review, the absence across ACT-R, SOAR, CLARION, LIDA, and BDI of any model of the multi-tiered self, any spiritually oriented reward, or any central ethical core [2], [3], [7]–[10], [13], remains a real and largely unexplored opening. Future work should build a working implementation, test the architecture's behavior with the Nafs and Taqwa Center switched on and off, sharpen how intention is actually encoded, and evaluate the system with real users over an extended period.

REFERENCES:

[1] J. Minninger and E. Dugan, Make Your Mind Work for You: New Mind Power Techniques to Improve Memory, Beat Procrastination, Increase Energy, and More, Hardcover, 1988.

[2] Misbah Yazdi, M. T., Akhlaq dar Qur'an [Ethics in the Qur'an], 3 vols., Qom: Imam Khomeini Educational and Research Institute, 1383 Sh. / 2004–2005.

[3] R. Hamani and M. Mokhtari, "Formulating an ideal model for the relationship between ethics and artificial intelligence: Emphasizing the ethical foundations of Allama Misbah Yazdi," Faslnāme-ye Akhlāq Pajuhī, vol. 8, no. 1, pp. 89–112, 2025, doi: 10.22034/Ethics.2025.51930.1777.

[4] Y. Wu, W. Guo, Z. Liu, H. Ji, Z. Xu, and D. Zhang, "How large language models encode theory-of-mind: A study on sparse parameter patterns," npj Artificial Intelligence, vol. 8, no. 1, 2025.

[5] A. Seddighi, S. A. Asghari, M. Romoozi, and H. Babaei, "Providing a trustworthy framework on the Social Internet of Vehicles based on fuzzy logic and game theory," International Journal of Multiphysics, vol. 18, no. 4, pp. 861–878, 2024.

[6] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in Proc. 34th Int. Conf. on Machine Learning (ICML 2017), 2017.

[7] J. R. Anderson, The Architecture of Cognition, Harvard University Press, 1983.

[8] J. E. Laird, A. Newell, and P. S. Rosenbloom, "SOAR: An architecture for general intelligence," Artificial Intelligence, vol. 33, no. 1, pp. 1–64, 1987.

[9] R. Sun, Ed., Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation, Cambridge University Press, 2006.

[10] S. Franklin, T. Madl, S. D'Mello, and J. Snaider, "LIDA: A systems-level architecture for cognition, emotion, and learning," IEEE Transactions on Autonomous Mental Development, vol. 6, no. 1, pp. 19–41, 2014.

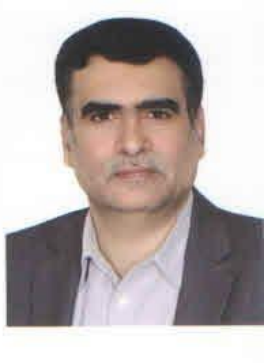
[11] W. van de Sanden, An Analysis of ACT-R and CLARION Representing Heuristic Strategies for Consumer Decision-Making: A Systematic Literature Review, Bachelor's thesis, Delft University of Technology, 2025.

[12] S. Kugele, "Constructivist procedural learning for grounded cognitive agents," Cognitive Systems Research, vol. 90, 101321, 2025.

[13] A. S. Rao and M. P. Georgeff, "BDI agents: From theory to practice," in Proc. First Int. Conf. on Multiagent Systems (ICMAS-95), pp. 312–319, 1995.

[14] A. Seddighi, S. A. Asghari, M. Romoozi, and H. Babaei, "A review of game theory-based trust modeling approaches in infrastructure-free social networks," Management Strategies and Engineering Sciences, vol. 7, no. 4, pp. 81–95, 2025.

How to cite: A. Khosravi, Z. Najafi, **RAVA: An Integrated Cognitive-Spiritual Architecture for Value-Aligned Decision-Making in Autonomous Intelligent Agents** , Journal of Distributed Computing and Systems (JDCS), Vol 9, Issue 1, Pages 111-126, 2026.



Arash Khosravi received his B.Sc. degree in Software Engineering from the University of Isfahan in 2003, his M.Sc. degree in Information Technology Engineering in 2013, and his Ph.D. degree in Information Technology Engineering, with a specialization in Information Systems, from Universiti Teknologi Malaysia in 2017. He is currently a faculty member at Mahallat Institute of Higher Education. His research focuses on Artificial Intelligence, particularly Natural Language Processing (NLP), Ontology Engineering, Large Language Models (LLMs), Knowledge Representation, Knowledge Graphs, and Neuro-Symbolic Artificial Intelligence. His recent work explores semantic architectures for intelligent systems, explainable and trustworthy AI, semantic interoperability, retrieval-augmented generation (RAG), and ontology-driven knowledge management. In addition to his current research interests, he has conducted research in business intelligence, recommender systems, customer knowledge management, data mining, text mining, and the application of information technology in healthcare. His ongoing research aims to bridge symbolic knowledge representation and data-driven AI through semantic frameworks that improve reasoning, explainability, and human-centered intelligent systems.

Email: khosravi.280@gmail.com

Zahra Najafi holds a Bachelor's degree in Computer Science from Mahallat Institute of Higher Education. She is actively engaged in research and practical applications within the fields of Ontology Engineering and Programming. Her research focuses on the design and development of ontologies, knowledge-based systems, data analysis, and the implementation of robust software solutions.Her research interests encompass the Internet of Things (IoT), Ontology and Knowledge Engineering, Knowledge Graphs, and the application of Artificial Intelligence in content processing and organization. Furthermore, she is dedicated to the development of intelligent data-driven systems and advanced Artificial Intelligence applications.

Email: z.najafi5761@gmail.com