How suited are cognitive architectures for implementing moral reasoning? – a Systematic Literature Review

Wojciech Hajdas[1]
Supervisor(s): Bernd Dudzik[1]**, Chenxu Hao**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Wojciech Hajdas
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Chenxu Hao, Catherine Oertel

Abstract

This paper surveys nine studies that implement aspects of moral reasoning within cognitive architectures (CAs) or CA-inspired frameworks. Its primary aim is to assess the viability of this approach for future research and to clarify the state of the domain. Two research paradigms emerge: (1) modeling human moral reasoning and (2) constructing artificial moral agents. Despite this distinction, all studies face similar challenges: fragmented reuse (each employs a different architecture), limited pre-programmed behaviors, and the absence of standardized benchmarks or metrics. Researchers remain optimistic about the explainability of their systems' behaviors and inner workings, yet often they acknowledge significant scalability and validation hurdles. Overall, CAs currently support only small-scale experiments; substantial further research – both empirical and into the theoretical basis of the field – is needed before these systems can attain real-world relevance.

1 Introduction

1.1 Background

Cognitive architectures (CAs) are computational frameworks designed to model and reproduce aspects of human cognition, serving as a foundation for the development of intelligent agents. Langley et al. [1] argue that, ideally, cognitive architectures should support belief maintenance, predictive simulation, planning, and decision execution, with each architecture being suited for a wide range of tasks.

Contrary to today's increasingly popular machine learning approaches, whose internal thinking and decision making are largely unexplainable and hidden in the form of "black boxes" [2], cognitive architectures offer better traceability in their decision-making process. This transparency is needed in ethically charged domains such as law, medicine, and governance, where stakeholders require clear accountability for decisions to uphold the legitimacy of these institutions and to build public trust with regard to these important digital decision-support systems [3]. Therefore, as argued by existing research [4], [2], [5], cognitive architectures offer a unique approach to enforcing AI ethics, as they enable the creation of systems capable of transparent thinking.

However, in order for systems to act ethically, they must be capable of moral reasoning. Moral reasoning, in short, is the process of evaluating actions as right or wrong [6]. Concepts of right and wrong are not easily defined; numerous ethical theories attempt to establish coherent systems of moral judgment based on different principles or underlying logics [7]. For practical purposes, many researchers equate an agent's ability to reason morally with its ability to follow a given ethical theory [8].

In addition to transparency of reasoning, another valuable quality in the design of moral agents (i.e., systems capable of autonomous moral reasoning) is similarity to human cognition. This is beneficial for several reasons, including enabling the system to interpret human behavior or to provide justifications that are understandable to humans [8].

In theory, cognitive architectures support both of these capabilities, making them promising candidates for the development of ethical systems.

1.2 Related Work

Related literature includes:

- ***Artificial Moral Agents: A Survey of the Current Status*** [9] - This literature review compiles and analyzes papers presenting Artificial Moral Agents. However, it does not focus on cognitive architectures. Only two studies analyzed in that review overlap with those included in this one.
- ***Cognitive Architectures for Artificial Intelligence Ethics*** [2] - This study presents a clear proposal and motivated call to action, encouraging researchers to consider cognitive architectures as tools for advancing AI ethics. While it references several studies also included in this review, it does not provide an in-depth analysis of them.
- ***The Case for Explicit Ethical Agents*** [8] - This work argues for the necessity of incorporating ethics into artificial agents. One of its key suggestions is the use of cognitive architectures to achieve this. It provides a broad overview of the field and briefly mentions some of the studies covered in this review.

All of these broader studies share the common goal of emphasizing the importance of developing moral agents. Two of them specifically propose cognitive architectures as a promising avenue. However, despite the fact that model implementations are crucial to the advancement of the field, no comprehensive review currently exists that systematically analyzes working systems.

This review seeks to address that gap by focusing specifically on implementations, with the aim of informing future researchers about possibilities, challenges and approaches before they commit to using CAs, or specific frameworks.

1.3 Research Question

Consequently, the main research question that defines the scope of this study is presented below.

How suited are cognitive architectures for implementing moral reasoning?

Suitability is defined here as a subjective, composite measure that considers the scale and capabilities of the implemented moral agents, the challenges encountered by researchers during development, and their overall attitudes toward the outcomes of their implementations.

In order to answer this broad and exploratory question in a structured and effective way, and to gather more information about the domain, the following sub-questions are proposed and elaborated on.

- **Which cognitive architectures, and in what configuration, are used in the implementation of moral reasoning?** - This question aims to outline the general technical landscape of the field, presenting the characteristics of the systems in which CAs are used, as well as the types of CAs involved.
- **How do researchers tackle moral reasoning in their studies?** - This question aims to present the theoretical basis of the ethical frameworks implemented and motivations behind them.

- **What are the results of the implementations?** - This question compiles the concrete decision making capabilities of the implemented systems and the conclusions drawn by researchers.
- **How do researchers who implemented these systems reflect on their work and envision its future** - This question captures researchers' perspectives on the systems they have created, highlights challenges and limitations researchers encountered, and supports them by gathering and synthesizing proposed directions for future work.

2 Methodology

This review is structured according to the PRISMA guidelines [10]. Some elements of the research process, such as the columns of the data extraction table or the initial scoping strategy were inspired by chapters of the book by Boland et al [11]. This section outlines the methodology of each research step. The entire review was conducted by a single researcher.

In Section 2.1 the selection criteria for the review are mentioned. Section 2.2 describes the strategy developed to find appropriate papers, while Section 2.3 presents results of the aforementioned search, along with a PRISMA-flow diagram [12]. In Section 2.4, data extraction methods are specified. Finally, Section 2.5 contains a description of the detailed methods used for data analysis and synthesis.

2.1 Paper eligibility

To identify relevant papers for the review, inclusion and exclusion criteria were created. They allow for the systematic selection of papers to be included in the data extraction.

Inclusion Criteria:

- The study describes an implemented system that deals with moral reasoning. That includes systems in which moral reasoning capabilities are only a part of the larger program.
- These programs must either be implemented using established cognitive architectures [13] or their frameworks must emulate human cognition - creating a new Cognitive Architecture.

Exclusion Criteria:

- Study uses Neural Network as part of their implementation.
- Paper is not written in English.

2.2 Search Strategy

Following initial scoping research, further guided by eligibility criteria and the research question, relevant domains and keywords were identified. The terms ***cognitive architectures*** and ***morals*** are considered essential for inclusion in this review. Actual queries are

expanded by ***ethics*** as a synonym for morals, as some papers don't differentiate between the two, and ***ACT-R, SOAR, LIDA*** as possible replacements for cognitive architectures, as some papers don't explicitly mention these systems as architectures. These three architectures are included by name due to their popularity and strong support communities [13]. Other architectures were excluded due to time constraints and scope limitations. The approximate query used and then adapted to all search engines is as follows:

```
("cognitive architectur*" OR "ACT-R" OR "SOAR" OR "LIDA") AND ("moral*" OR
                                "ethic*")
```

The `*` characters represent any possible string of characters. Keyword searches were applied at the highest level supported by each database (full-text when available; otherwise, abstract, title, and keywords). Full queries per database are recorded in Appendix C.

2.2.1 Search Engines

Five search engines were used: *IEEE Xplore* [14], *Scopus* [15], *Web of Science* [16], *ACM Digital Library* [17], and *SpringerLink* [18]. All were chosen due to their prevalence in the initial internet scoping searches. Specific queries for each database are provided in the Appendix A. Additionally, eight papers were identified on other websites, such as *academia.edu* [19], during the initial scoping searches.

2.3 Search Results

After executing the queries, the titles of retrieved papers were screened, and irrelevant studies were excluded.. The rest of the papers were saved into Zotero citation management software [20]. Filtering continued by screening abstracts and applying the eligibility criteria. Finally, papers were screened by their full text. The entire process, including the specific number of results, is presented in the PRISMA flow diagram in Figure 1.

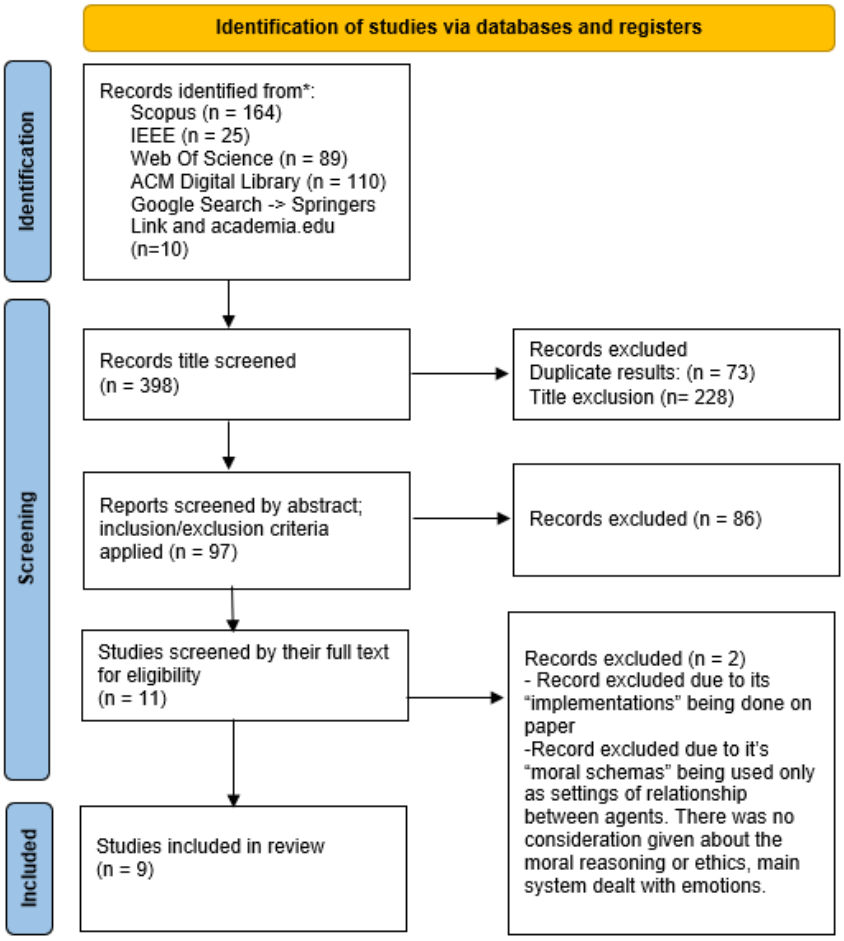

Figure 1: Adapted PRISMA flowchart.

Many papers were excluded during title screening, as the keyword combination "soar" and "morals" often returned results from humanities studies.

2.4 Data extraction methods

To answer the four sub-questions effectively, more specific "extraction questions" were formulated for each. Their main purpose is to guide the data extraction process in a systematic and replicable manner. Table 1 presents the proposed extraction questions alongside the sub-questions they support.

Table 1: Overview of the extraction questions

Sub-Questions	**Extraction Questions**
1.Which cognitive architectures, and in what configuration, are used in the implementation of moral reasoning?	1.1 What cognitive architectures are used? 1.2 How are they configured with other system elements (sensors, GUI, etc.)? 1.3 What motivates these choices?
2. How do researchers tackle moral reasoning in their studies?	2.1 What are the ethical theories or assumptions on which researchers focus and base their implementations on? 2.2 What motivates these approaches?
3. What are the results of the implementations?	3.1 What specific actions of moral/ethical weight can the systems enact? 3.2 What evaluation methods/metrics are used? 3.3 What conclusions were drawn?
4. How do researchers reflect on their work and its future?	4.1 What concerns and discussion points are raised? 4.2 What are the future work recommendations?

For each study that passed the selection process, a new entry was created in an MS Excel data extraction table. Each entry consists of two rows, with the extraction questions acting as columns, plus one additional cell. The first row was filled manually by the researcher, extracting quotes from the text that answered the corresponding questions. The additional cell was used for notes or contextual quotes relevant to the later analysis phase. The method of filling the second row, using AI tools, is specified in the next sub-section 2.4.1.

2.4.1 Use of AI in data extraction

In order to speed up the data extraction process as well as make it more robust, a Large Language Model program was used. I decided on the use of Google's Notebook LM [21] due to its verifiability, as each answer given by the system is supported by a highlighted quote from the source paper, allowing for easy manual correctness check. Notebook LM was utilized by uploading the PDF of each study from the final selection in separate sessions - "Notebooks" and giving it prompts for each extraction question:

- **"Answer this question shortly based on the implementation in the text:"** + the question's text - for extraction questions 1.1 - 3.2
- **"Answer this question shortly based on the text:"** + the question's text - for extraction questions 3.3 - 4.2

These prompts were developed by a couple of exploratory queries, after which I deducted that in some cases Google LM would not base its answers on a proposed implementation, but on, for example, considerations authors give to other approaches. Thus, questions that were based on the implementation required the additional specification.

Relevant parts of each answer, after manual verification, were saved into the aforementioned second row of the data extraction entry, either as highlighted quotes or summaries.

In many cases this process allowed for the identification of relevant quotes that were omitted during manual search, which was especially true for studies of longer length.

2.5 Methods of analysis

Based on the entries in the data extraction table, an answer for each sub-question will be given, supported by a concise table containing relevant data-extraction table fields. The main research question will be addressed in Sections; Discussion 5 and Conclusion 6, by synthesizing the sub-question answers, along with relevant insights or groupings identified during the data analysis process.

3 Results

In this section results of the data extraction and synthesis process are presented. Papers that were taken into account during the extraction process are [22], [23], [24], [25], [26], [27], [28], [29], [30].

Next, Sections from 3.1 to 3.4 present compiled answers to the research sub-questions.

Before any in depth analysis is conducted and presented, it is important to note that all of the presented papers fall within two groups;

- (1) containing papers that try to model human moral reasoning using CAs
- (2) containing papers focused on creating artificial cognitive moral agents with the use of CAs

When the goal is to model human cognition (e.g., [24], [27], [29]), researchers often accept non-optimal moral behavior as long as it corresponds to human variability. Conversely, in studies that aim to implement moral reasoning in robots (e.g., [22], [23], [25], [26], [28], [30]), the focus shifts toward developing systems that strive for ideal moral performance, according to various definitions. Analyzing studies while having these divergent goals in mind, provides more meaningful insights. Papers from both of the groups, however, leverage cognitive abilities of the CAs and deal with similar considerations and domains, which is why both of the groups are included in this review.

Complete results for each paper are presented in the data extraction tables provided in the appendices. Appendix A contains the table for the human modeling group, while Appendix B contains the data for the artificial moral agents group. Data extraction entries were created by summarizing the relevant quotes identified in the studies. None of the conclusions or factual statements made by the original researchers were independently verified as part of this work. In some cases, additional data was available, but information deemed less central to this domain was omitted due to length constraints.

3.1 Technical Choices in Implementations

3.1.1 Choices of CAs and Motivations Behind Them

Despite the often similar qualities that researchers seek in their choice of CAs, each study arrives at distinct conclusions, with no CA repeated across the selected papers; however, some overarching goals of the implementations show partial overlap within certain subgroups. For example:

- [22] and [23], both aim to develop autonomous moral robots;
- [27] and [29] - studies from the human moral reasoning group, both emphasize the lack of emotional considerations in existing CAs as motivation for proposing new frameworks;
- [25] and [26], both present patient-care moral agents based on constraint-obedience;
- [24], [26], [28], all strive to allow their agents to make premeditated decisions by using specific mechanisms within CAs: [24] - highlights Clarion's [31] focus on motivation, [28] - mentions ARCADIA's [32] focus on intentional action and [26] - cites SOAR's [33] ability to make informed decisions.

Importantly, in all reviewed studies, the need to address a specific moral behavior or research question preceded the choice of architecture. None of the papers primarily aimed to evaluate a CA's capability for moral reasoning in a comparative or exploratory sense. Rather, the CAs were selected as tools believed to support the researchers' specific objectives.

A detailed analysis of how particular architectural mechanisms influence researchers' choices is beyond the scope of this study, as it would require a deep technical understanding of each CA, which are complex systems in their own right. For example, in study [22], ACT-R [34] was chosen for its "spreading activation modules" and "knowledge processing mechanisms, based on the interaction between short-term and long-term memory", being able to draw in-depth conclusions based on this information would require longer period of research. The more specific motivations provided by the authors are documented in the appendices A and B.

3.1.2 Use of CAs within Larger Systems

Most of the presented systems ([24], [27], [26], [29]) are used only as computer simulations, presenting results within the kernels. A subset of studies ([25], [28], [30]) employ simple visual interfaces to represent the agents' decision environments, primarily to aid researcher interpretation and debugging. Only two studies, [22] and [23], integrate their systems more fully into real-world contexts. [22] utilizes a conversational, animated graphical interface, while [23] implements the architecture in a physical robot designed to operate autonomously in human environments.

3.2 Focus areas in Moral Reasoning domain -SQ2

3.2.1 Choices of Ethical Theories for Implementations

Researchers differ in their emphasis on particular ethical theories or morality-related cognitive processes. However, nearly all papers, except [27] and [29], explicitly engage with the ethical distinction between consequentialist/utilitarian and deontological approaches. Some studies also mention virtue ethics as a third main theory. Below short definitions of these leading theories are presented:

- Utilitarianism - states that the right action is the one that maximizes happiness for the most amount of people. Actions are judged based on their consequences [35].
- Deontological Ethics - states that actions are considered morally good not based on their consequences but on the moral laws that they obey [36].
- Virtue Ethics - primarily concerns itself with the character traits that are essential for morally good action. It cannot be easily classified on the utilitarian-deontological spectrum [37].

Most studies ([23], [25], [27], [30]) feature agents governed by deontological rules and constraints. One study, [22], explores a virtue ethics inspired agent, while another, [26], implements a utilitarian decision-making model. Two papers, [24] and [30], contrast agents employing different ethical paradigms within the same experimental setting.

In addition to developing agents that exhibit moral behavior, some studies also aim to simulate specific cognitive or psychological phenomena within moral contexts. [28], for example, models artificial self-control, while [29] presents a system capable of acquiring new moral norms through learning mechanisms.

3.2.2 Motivations behind Choices of Ethical Theories

The choice of ethical frameworks is closely tied to each study's overarching research aims. For instance, [23], [26], and [25] justify their ethical models by their relevance to real-world, socially embedded robot behavior. In contrast, studies focused on human moral reasoning ([24], [27], [29]) anchor their choices in theoretical and empirical considerations aimed at achieving psychological plausibility. The remaining studies prioritize the exploration of under-examined phenomena or conceptual gaps in the field.

Only one study [24] acknowledged the possible impact of culture on their choices of ethical theories.

3.3 Results of the Implementations

3.3.1 Agent capabilities

The implemented agents across all studies exhibit relatively limited sets of morally consequential actions. Due to the high variability of implementation contexts and experimental designs, these capabilities do not allow for easy systematic categorization. However, a few common trends can be identified:

- Some agents are capable of disobeying human commands on moral grounds - [23], [25], and [26];
- Others are designed to judge explicit moral dilemmas, such as variations of the trolley problem - [24] and [30].

The remaining implementations involve agents that perform highly context-specific moral actions, typically chosen from limited, hard-coded options. For instance, in [28], the agent's only decision involves selecting which of two colored keys to retrieve and return to base within a simulated environment. Similarly, in [27], the agent can choose between four predefined acts of interaction with other agents. Due to this specificity and variability, data entries in the Appendices A and B provide the best representation of the topic.

3.3.2 Validation methods

Notably, none of the studies apply standardized benchmarks or shared evaluation criteria for validating their models. In fact, several studies, including [23], [25], [26], and [28] - do not describe any formal validation procedures. Instead, they only provide actions which agents should take for a specific run to be considered successful.

All studies aiming to model human moral reasoning compare agent performance to human data. In [24] and [27], comparisons are made quantitatively, with experimental results from human participants directly contrasted with those of the artificial agents. [29] validates its approach by aligning outcomes with findings from neuro-cognitive literature.

Among the studies focused on artificial morality, only [22] includes a dedicated validation process. This involved an empirical study with 64 participants, in which the perceived ethicality, persuasiveness, and engagement of the conversational agent were evaluated.

3.3.3 Conclusions in the studies

The conclusions drawn by the researchers across all studies are generally positive and optimistic. Several common trends can be identified:

- All studies from the human modeling group ([24], [27], [29]) conclude that their chosen cognitive architectures are effective for modeling aspects of human moral reasoning.
- Studies [25] and [28] acknowledge the limited scale of their implementations but highlight the potential of the selected architectures for further development.
- Studies [22] and [23] report that the integration of their selected moral theories improved the performance and quality of their agents.
- Studies [26] and [30] state the facts of their implementations as the main conclusions derived from their research.

3.4 Reflections and Future Directions

3.4.1 Discussions and Concerns

Several studies explicitly reflect on the limitations of their implementations and raise concerns about the quality and scope of their experiments. Notable concerns include:

- [30] stated that testing of autonomous systems is still in its infancy, thus testing it's agents is challenging. [22] also noted problems with validating its solution due to a low number of validation study participants.
- [29] and [28] acknowledge limitations in their chosen architectures and emphasize that the tasks modeled in their experiments were highly simplified and hard-coded, lacking dynamic problem detection or adaptability.

Additionally, both [22] and [23] when reflecting on the broader implications of their work, emphasize the importance of embedding ethical reasoning in future robotic systems.

Studies [25], [26], and [27] didn't provide any detailed discussion.

3.4.2 Proposals for Future Research

While most suggestions for future research are study-specific, several recurring themes can be identified:

- [23], [24], and [28] propose deeper integration of their moral reasoning systems with additional mechanisms already present within their respective CAs.
- [25] and [30] suggest enhancing their chosen architectures with capabilities currently unavailable in the frameworks they use.
- [22] and [27] highlight the need for more extensive testing to validate their systems.

Many of the studies also include additional, context-specific suggestions for future research, which are documented in Appendices A and B.

4 Responsible Research

In this Section ethicality and reproducibility of the research will be discussed.

4.1 Risks during Data Extraction and Presentation

This study was conducted by a single researcher over a relatively short time frame of ten weeks. This limitation introduces several potential risks, particularly regarding the subjectivity of the data extraction process. Many of the reviewed studies did not provide clear or direct answers to the predefined extraction questions, requiring interpretation across various parts of the texts. This interpretive process is inherently subjective and could have yielded different results if conducted by another researcher.

Risks also extend to the presentation of results. The identification and emphasis of specific trends derived from the texts are based on subjective interpretation. Different researchers

may have highlighted different aspects or drawn alternative conclusions from the same data.

Additionally, lack of specialist knowledge concerning philosophy, neuro-cognition or ethics, might have affected drawn conclusions, discussions and focus within data presentation. These risks are not evenly distributed however, as studies present different levels of theoretical ethical complication.

To mitigate these risks, redacted versions of the data extraction tables are included in the appendices, along with concise summaries of each study. These supplementary materials are intended to provide transparency and allow readers to assess the consistency and reliability of the reported findings.

However, it must be acknowledged that concerns regarding the reliability of the presented extraction tables should be taken into account. A comprehensive, quote-based data extraction table was created but ultimately not included, as it would have been excessively large and impractical for meaningful analysis without significant time investment.

4.2 Risks during Paper Selection

Risks of bias and limited reproducibility during the paper selection process were mitigated through a clearly described search strategy and the application of PRISMA [10] guidelines. Nevertheless, certain limitations remain due to the study being conducted by a single researcher. Notably, there is a risk of overlooking relevant literature, as the scope and depth of the search process were necessarily constrained by time and available resources.

The choice of the inclusion of non-peer reviewed, conference papers in the selection process, with one such paper taking part in data extraction - [26] - may also raise some quality of data concerns. Despite this, due to the relative unpopularity of the field, any paper was deemed to have the potential of increasing the level of current knowledge.

Additionally, excluding papers not written in English might have excluded relevant studies, especially ones based on ethical theories prevalent out of Anglo-sphere. Negating this risk was deemed impossible, given the scale of the review.

4.3 Use of LLMs

LLMs were used during two phases of the creation of this paper. First, as mentioned previously, Google LM was used as a data-scouting tool to facilitate data extraction by highlighting and compiling relevant quotes. To mitigate the risks associated with incorrect or fabricated output, each quote was thoroughly checked for accuracy before being added to the data extraction table.

The second use of LLMs occurred during the editing process. ChatGPT [38] was employed as a proofreading and editing tool to improve the grammar and clarity of sentences. It is important to note that ChatGPT was not used for writing new content, with the exception of abstract, which was in turn, edited manually.

5 Discussion

Systematic review of nine studies that implement moral reasoning reveals several patterns that are present within the current state of the field and its future.

5.1 Fragmentation and Diversity in the Field

One of the most significant findings is the lack of repetition in the cognitive architectures (CAs) chosen for implementation. This diversity contrasts sharply with the broader field of artificial intelligence, where certain solutions eventually emerge as standard practice, allowing research to build incrementally on previous work [39]. This phenomenon displays the relative infancy of the field, with researchers pursuing fundamentally different approaches to address similar challenges within the domain.

The absence of clear research trajectories also raises questions about the motivations behind selecting specific CAs. Without a deeper comparative analysis of the field, it is difficult to determine whether CAs were chosen for their particular strengths or simply because of researchers' personal preferences or prior experience.

Moreover, this diversity is not limited to the choice of CAs. Even though most studies acknowledge central ethical dilemmas, their modeled ethical theories and behaviors do not follow any consistent trends. This may reflect the inherent complexity of ethics as a discipline, emphasizing the need for developing a scientific consensus and practical guidelines for the behavior of ethical machines.

Finally, this fragmentation is evident in the evaluation methods across the studies. The lack of shared benchmarks, standards, or quantitative validation, combined with the absence of universally accepted metrics, makes direct comparisons between approaches highly subjective and potentially unreliable. This issue reinforces the need for a stronger theoretical foundation and more rigorous practical standards, especially as some researchers have already noted that autonomous system testing is still in its early stages.

5.2 Two Research Paradigms

The distinction between studies focused on modeling human moral reasoning and those aimed at developing purely artificial moral agents reflects a fundamental divide in research objectives. This divide also raises important questions about what constitutes meaningful progress in the field.

This divide naturally prompts further discussion about the future design of artificial moral agents. On one hand, agents that emulate human cognition may offer greater transparency and interpretability, as well as improved predictive capacity for understanding human behavior - although they also risk inheriting human biases and cognitive limitations. On the other hand, agents that implement purely artificial moral reasoning could provide more objective and impartial judgments, but may appear alien, less intuitive, or even less trustworthy to human collaborators.

Interestingly, despite these differing research goals, (human-centered models striving to align with empirical data, and models striving for idealized, formal ethics), both groups face similar challenges when using cognitive architectures. In particular, both struggle with small-scale, highly controlled experiments and unrealistic scenarios that limit the generalizability of their findings. This overlap in challenges suggests that the fundamental difficulties involved in implementing robust moral reasoning have not yet been resolved, and that greater collaboration between the two research trends could help overcome these obstacles.

5.3 Limited Scope and Scalability Challenges

Most of the reviewed studies demonstrate capabilities that are highly limited in scope, often restricted to unrealistic and highly bounded scenarios. Many of these systems – particularly those developed in the artificial agents group – do not improve upon existing solutions to similar problems that do not utilize cognitive architectures.

This issue, like the fragmentation noted earlier in Section 5.1, may stem from the inherent complexity of moral reasoning. Real-world moral decisions often involve competing values and uncertain consequences that are difficult to formalize in a consistent and reliable way. The trolley problem, while philosophically interesting, bears little resemblance to the nuanced moral challenges encountered in practice, such as those arising in healthcare, autonomous driving, or social media content moderation.

Another factor limiting the capabilities of these implementations is the steep learning curve and significant commitment required to use cognitive architectures effectively. Integrating them with other system components introduces additional complexity. This may explain why many researchers focus on deontological constraints - i.e. restricting agents from certain actions, as this approach is not only seen as useful but is also comparatively easier to implement than the more sophisticated models of ethical reasoning.

5.4 Analysis of domains related to the Research Question

The main research question "How suited are cognitive architectures for implementing moral reasoning?" requires analysis of its separate components, given below. A concise answer will be included in the Conclusion – Section 6, to avoid redundancy.

5.4.1 Scale and Capabilities

Despite their limited capabilities, as discussed earlier, cognitive architectures (CAs) offer structured reasoning and explicit knowledge representation that can support transparent moral decision-making. They can also reproduce subtle behaviors observed in human data. However, while some CAs have built-in cognitive features, only one study demonstrated adaptation of a system's moral framework over time [29]. Most other experiments relied on fixed, hard-coded moral theories.

The systems perform well and can display their strengths on simple, unrealistic scenarios, but are not yet proven on more complicated, realistic tasks. This limitation reflects a broader

challenge in the field, where the complexity of real-world moral reasoning remains difficult to capture within the bounded environments that current implementations can handle.

5.4.2 Development Challenges

Several identified problems affect the development process. The fact that no two studies employed the same architecture indicates significant learning curves and limited knowledge transferability between implementations. Theoretical challenges primarily comprise the lack of validation methods for different moral theories. Additionally, the complexity of CAs combined with the complexity of the field of ethics potentially increases the barrier of entry for researchers without extensive cognitive science or philosophical backgrounds.

However, efforts to address these problems are already underway, with one study [30] presenting a platform designed for easier moral agent creation. This development suggests recognition within the research community of the need for more accessible tools and frameworks to advance the field.

5.4.3 Researcher Attitudes

Despite the identified challenges, researcher attitudes remain positive across all studies. This optimism appears well-founded for human moral reasoning modeling, where cognitive architectures perform well in replicating human behavior. For artificial moral agents, attitudes are more cautious but still generally positive, with researchers viewing current limitations as natural to an emerging field of science, (based on the future work recommendations often focusing on development of CAs), rather than fundamental barriers.

6 Conclusions and Future Work

6.1 Conclusions

This literature review analyzed nine studies that present systems capable of moral reasoning implemented using cognitive architectures. From each paper, data were extracted to address the main research question: **How suited are cognitive architectures for implementing moral reasoning?** This was supported by sub-questions concerning technical configurations, approaches to moral reasoning, the results of the studies, and the researchers' attitudes.

6.1.1 Key Findings

- In none of the studies was the choice of CA repeated. This indicates either diversity in optimal solutions (reflecting the diversity of cognitive architecture characteristics) or a lack of systematic analysis that could identify superior approaches.
- The research field is progressing under two primary paradigms: modeling human moral reasoning and creating cognitive artificial moral agents. This division influences what is considered a successful implementation and how progress is measured. In particular, studies focused on human modeling may accept suboptimal moral behavior if it matches human data. Despite these differences, both paradigms face similar implementation challenges.

- The scope and scale of current implementations are severely limited. Existing systems typically operate within highly constrained, unrealistic simulations or allow very limited freedom of choice. This limitation reflects both technical constraints involved in scaling cognitive architectures and theoretical challenges in formalizing complex moral reasoning.
- The lack of standardized evaluation methods and shared benchmarks slows progress in the field by making it difficult to compare different approaches and recognize which solutions work best.
- Cognitive architectures display desired behaviors and characteristics in limited simulation environments, including explainability of decisions or observability of the reasoning process.

6.1.2 Answer to the Main Research Question

Considering all of the limitations, challenges as well as positive traits and optimistic attitudes of researchers, an answer to the main research question can be formulated.

Cognitive architectures are conditionally suitable for modeling of moral reasoning. They exhibit their key strengths (transparency and explainability) even in current limited implementations, and considering researchers' generally optimistic attitudes, continued research in this domain is warranted. However, significant advances in scalability and standardization are needed before definitive conclusions about real-world applicability can be drawn.

Cognitive Architectures are a complex tool, used for implementing of moral reasoning – also a complex task. This combination might be the main cause of the rather unimpressive results, even despite the validity of the approach.

6.2 Future Work

Based on this systematic review, several research directions emerge as priorities:

- Systematic Literature Review analyzing proposals for the ethical guidelines of autonomous systems.
- Development of benchmarks and standards that could be used in evaluation of different moral agents and ethical theories.
- Comparative studies systematically evaluating different CAs on identical moral reasoning tasks.

A Human Moral Reasoning Group Redacted Data Extraction Table

Sheet intentionally left blank. Table starts at the next page.

Table 2: Extraction table for human modeling group Part 1

Questions/Papers	**Two Models of Moral Judgment** [24]	**Emotional biologically inspired cognitive architecture** [27]	**Neurocomputational model of moral behavior** [29]
What cognitive architectures are used?	Clarion	eBICA - emotional Biologically Inspired cognitive architecture - proposed in this study.	MONE -MOral Neural Engine
In what configurations with other elements of the systems are they used? (sensors, user input, GUI, avatars etc.)	No notable additional elements. The system is a kernel simulation, outputting results in a numerical form with simulated reaction time data.	No visualization, text based input and output.	Visual inputs that are interpreted by MONE. No other elements mentioned.
What are the motivations behind using these specific cognitive architectures and their configurations?	Clarion accounts for modeling basis human motivations which influence further behaviour, this emphasis helps in explaining the integration of cognitive systems with motivational considerations. Additionally, Clarion focuses on the broad cognition-motivation-environment interaction, in contrary to narrow focus of some architectures only on cognition. According to researchers those 2 qualities help with explaining the human moral judgment.	Filling the research gap of including emotions into cognitive systems. Lack of emotions could be the problematic for the acceptance of virtual systems as equal partners. (Moral reasoning is seen as a important part of emotional system.)	The goal is to replicate, as faithfully as possible, the structure that in the brain gives rise to moral cognition. Visual and "taste" inputs enable simulating many moral situations embedded in a microworld. MONE offers emotional approach to the morality.
What ethical theories or assumptions do researchers focus on and base their implementations on?	Two simulation models reflect different assumptions: Model 1 applies an emotion-reason conflict theory, assuming deontological (emotion-based) decisions are faster and higher-level, whereas utilitarian calculations are slower and more deliberate. This is implemented using Clarion's action and rational subsystems. Model 2 relies on a motivationally based moral judgment theory derived directly from Clarion. It rejects a strict emotion-reason division, proposing that implicit and explicit processes interact, influenced by individual motivations.	The system uses moral schemas representing higher-order appraisals, which establish "normal" values for self-related opinions. These schemas influence action probabilities to correct deviations and constrain emotional system actions. This is another form of constraint-based ethics.	The research assumes moral cognition is primarily emotional, emphasizing the development of guilt and shame as drivers of moral thinking. Morality is viewed as inherently social, formed by positive and negative reactions from group members.
What are the motivations behind the specific approaches to ethics?	The approach is motivated by a desire to understand how different ethical theories influence agent behavior and decision-making speed, particularly contrasting fast, emotion-based decisions with slower, reasoned ones.	The use of moral schemas is motivated by the aim to better model human emotional intelligence and explain social behaviors such as maintaining roles or reinforcing hierarchies. These schemas improve alignment with observed human interactions.	The approach assumes: - Morality is a learned emotional process, - It does not originate from a single mechanism, - It can demonstrate how moral norms emerge from interactions and reinforcement.

Table 3: Extraction table for human modeling group Part 2

Questions/Papers	**Two Models of Moral Judgment** [24]	**Emotional biologically inspired cognitive architecture** [27]	**Neurocomputational model of moral behavior** [29]
What specific actions of moral/ethical weight can the systems enact?	In simulation, agents rate the moral permissibility of actions on a Likert scale, particularly in variants of the trolley problem. Examples include evaluating the morality of pushing a person onto the tracks versus dropping a person onto tracks by pressing a button to redirect a train and then capturing graded moral judgments.	In a simulated social environment, agents choose among four interaction types - hit, yield, greet, or ignore - to manage their status in a social hierarchy. Moral schemas are used to constrain emotionally driven choices, reinforcing or moderating behaviors based on internalized norms.	The model can choose to collect and eat an object or abstain, based on learned consequences of past actions. It has the ability to acquire and internalize moral norms over time, leading to increasingly ethical behavior.
What evaluation methods and metrics are used?	The models are evaluated by comparing action selection speed and Likert scale moral ratings with human performance data, assessing alignment with human moral judgments.	The system's performance is assessed through analysis between two groups: - all virtual agents group and - virtual agents and human participants group. Statistical and qualitative analyses are used to assess whether the implemented moral schemas can reproduce human-like social behaviors, such as hierarchy enforcement.	The evaluation involves three successive simulation runs, each introducing new variables related to the agent's internal state (e.g., hunger). Performance is measured by tracking how many times the agent chooses to eat an apple that does not belong to it, assessing moral learning across runs.
What are the main conclusions derived by the researchers?	In the comparison of models: 1. Model 1 was able to partially replicate human judgment patterns, but failed to match data for impersonal dilemmas. 2. Model 2 (motivationally driven) successfully captured all major aspects of the human data. Researchers concluded that motivational dynamics may be better basis for moral judgment than the simple conflict of emotion vs. reason, making Clarion a suitable architecture for modeling human moral cognition.	Researchers concluded that by using moral schemas capable of overriding some emotionally driven behaviors, the eBICA framework can replicate human-like social behaviors, such as maintaining social hierarchies. However, no specific performance figures were provided.	The study using MONE concluded that intact moral reasoning requires an intact amygdala, and effective decision-making that incorporates both internal drives and moral norms depends on the ventromedial prefrontal cortex. These findings are consistent with existing neurocognitive human data. Researchers also noted that MONE fills a gap, as prior neurocomputational models of moral behavior were lacking.
What are the main concerns and points brought up in discussions?	The discussion engaged with alternative theories of moral reasoning, suggesting other possible frameworks that could inform the implementation of ethical systems.	None identified.	The MONE model has several limitations: 1. It only simulates a single type of moral scenario (stealing). 2. It does not engage with cognitively grounded theories that view moral decisions as deliberative rather than emotional. 2. The experimental setting lacks realism, which weakens validity.
What are the propositions in future work recommendations?	Recommendations for Clarion include: 1.Integrating a wider range of data into the system's design and evaluation. 2. Activating and utilizing Clarion's more complex subsystems. 3. Exploring drives and goals relevant to everyday moral contexts. 3.Adapting the architecture to account for cultural influences on morality.	Future work for the eBICA framework includes: 1. Continuing simulations to test whether moral schemas can match human behavior in small groups. 2. More testing in larger groups that accounts for varied social phenomena. 3. Emphasizing more comprehensive testing and evaluation.	None proposed.

B Appendix B: Cognitive Artificial Agents Data Extraction Table

Sheet intentionally left blank. Table starts at the next page.

Table 4: Extraction table for cognitive Artificial Agent group Part 1

Questions/Papers	**A Storytelling Robot Managing Persuasive and Ethical Stances via ACT-R: An Exploratory Study** [22]	**An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures** [23]	**Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture** [25]
What cognitive architectures are used?	ACT-R	The Distributed, Integrated, Affect, Reflection, Cognitive (DIARC) Robot Architecture	LIDA
In what configurations with other elements of the systems are they used? (sensors, user input, GUI, avatars etc.)	System uses a conversational agent GUI interface that connects to ACT-R kernel. The kernel processes user input in real time.	The robot contains a LIDAR sensor for 3D data about environment, laser sensor for navigation as well as, place recognition and natural language understanding modules.	LIDA agents operates in an internal simulation environment with "human patients" inside. Diagnostic Panel for the simulation is available.
What are the motivations behind using these specific cognitive architectures and their configurations?	ACT-R was used for its spreading activation modules that retrieve and activate rules related to dialogue management, enabling autonomous behavior in non-sequential environments. Additionally, its knowledge processing mechanisms, based on the interaction between short-term and long-term memory, provide the agent with flexible and adaptable dialogue strategies.	As researchers argue, unlike other classic cognitive architectures, DIARC's polylithic nature is designed to enable autonomous, long term robotic operation. The sensors are supposed to provide additional context that influences the moral norms in place. DIARC with the combination with the sensors enables more detailed and specific command refusals. This explainability enhances trust.	LIDA is one of the few cognitive models which are neuroscientifically plausible and provides a plausible account for functional consciousness, attention, feelings, and emotions and has been partially implemented.
What ethical theories or assumptions do researchers focus on and base their implementations on?	The system adopts Aristotelian virtue ethics.	Researchers implement a context-specific deontological approach, where recommended or forbidden actions vary depending on the environment. This is a constraint-based ethical system, where violating a constraint is considered immoral.	Ethical system used is behaviour limited by simple constraints. No complicated reasoning processes are used.
What are the motivations behind the specific approaches to ethics?	The Virtue Ethics approach is motivated by Aristotle's view that cultivating a virtuous character, specifically open-mindedness, enhances the persuasiveness of arguments. This is further supported by Virtue Argumentation Theory (VAT). Additionally, the approach aims to fill a research gap, based on exploratory studies suggesting that combining persuasive storytelling with open-mindedness can effectively encourage users to reconsider their prior beliefs.	The motivation stems from the need to ensure that social robots are accepted in human environments by aligning with human moral and social norms. Key goals include: 1.Providing robust explanations for refusals to avoid misleading humans or undermining their moral expectations. 2.Enabling cooperation in shared environments.	Simpler ethical models are used due to practical limitations: implementing complex theories like utilitarianism or hybrid top-down/bottom-up approaches is too difficult, especially when actions affect multiple humans and lead to computational or design challenges.

Table 5: Extraction table for cognitive Artificial Agent group Part 2

Questions/Papers	**A Storytelling Robot Managing Persuasive and Ethical Stances via ACT-R: An Exploratory Study** [22]	**An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures** [23]	**Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture** [25]
What specific actions of moral/ethical weight can the systems enact?	The conversational agent selects from predetermined dialogue options based on user input, applying virtue ethics to adapt and justify moral reasoning in real time. For example, the agent refrains from insisting on mask-wearing when the user reports an allergy-demonstrating context-sensitive moral consideration.	A robot tasked with enforcing COVID-19 social distancing enters rooms to photograph occupants, except in morally sensitive areas like bathrooms, where such actions are deemed unethical. The robot is capable of refusing commands and providing moral justification for non-compliance.	Agents can make situational decisions, such as whether to complete feeding a patient when an urgent or non-urgent call is heard. These are classified as simple but morally relevant actions based on contextual awareness.
What evaluation methods and metrics are used?	A small-scale user study involving 64 participants was conducted. Participants interacted with the robot and completed questionnaires measuring perceived ethicality, persuasiveness, and engagement.	The system is evaluated based on behavior checks: 1. If the robot refuses to take a picture in the washroom, the action is considered ethically successful. 2. The DIARC architecture is assessed by analyzing robot logs and observing whether it can propose plans for action or justify disobedience.	No specific evaluation method is reported beyond the existence of a simulation.
What are the main conclusions derived by the researchers?	Integrating virtue ethics into a persuasive robot led to an increased tendency among users to re-evaluate their previously held beliefs. However, no definitive conclusion was drawn regarding the effectiveness of ACT-R as a framework for this purpose.	Researchers stated that they successfully developed a norm-aware task planner to achieve context-sensitive moral cognition in robots.	The researchers concluded that full ethical frameworks are currently too complex to implement, but simplified, constrained ethical decision-making is feasible using current technologies. Cognitive architectures can play a significant role in developing such systems, as they already model many mechanisms relevant to ethical reasoning.
What are the main concerns and points brought up in discussions?	A small number of participants limited the study's ability to generalize results. Some users observed inconsistencies between the robot's narrative style and moral stance. Additionally, researchers emphasized on "Ethics by Design" approach-embedding ethical considerations from the outset.	Researchers stressed that robots must be sensitive to human moral norms, which are often dynamic and highly context-dependent.	Discussion points were restated as conclusions, without a separate section or detailed analysis of limitations or challenges.
What are the propositions in future work recommendations?	Researchers propose: 1. Extending the system to a real physical robot to evaluate its performance in real-world conditions. 2. Testing with a larger and more diverse participant group to improve generalizability. 3. Increasing the robot's autonomy and enhancing its ability to recognize and respond to a wider range of social contexts.	Future work includes: 1. Generating natural language refusals that express context-sensitive politeness. 2. Developing the robot's ability for social navigation, particularly in group interactions. 3.Incorporating more DIARC components for norm analysis and task planning, expanding its ethical reasoning capabilities.	For the LIDA architecture, proposed improvements include: 1.Implementing more complex top-down and bottom-up moral judgment mechanisms. 2.Completing the integration of volitional decision-making. 3.Addressing metacognitive-level constraints, which currently exceed LIDA's capabilities.

Table 6: Extraction table for cognitive Artificial Agent group Part 3

Questions/Papers	**Application of Soar Cognitive Agent Based on Utilitarian Ethics Theory for Home Service Robots** [26]	**Self-control on the path toward artificial moral agency** [28]	**A Cognitive Architecture for Verifiable System Ethics via Explainable Autonomy** [30]
What cognitive architectures are used?	SOAR	ARCADIA	Agents are built on the Cogent platform.
In what configurations with other elements of the systems are they used? (sensors, user input, GUI, avatars etc.)	The system contains a Ubuntu kernel based interface with user health data provided with each run.	Simple graphical simulation based on Minigrid1 with simplistic GUI.	Cogent provides a visualization of the agent, which is intended to help the programmers to develop the agent, supporting their interactive and incremental development.
What are the motivations behind using these specific cognitive architectures and their configurations?	During previous work researchers identified SOAR as a cognitive agent aiming at human-level thinking which can make informed decisions to solve problems.	One of the primary phenomena ARCADIA was conceived to account for is intentional action - key to self control and thus, moral action. It was also chosen due to it's simple design that makes very few theoretical commitments.	A cogent is specified using a Domain Specific Language (DSL), which provides high-level abstract features based on the theory of cognitive coherence [18]. Its features aim to facilitate explicit specification of the perception and deliberation (reasoning) mechanisms. Allowing users to express deliberation models could help them avoid steep learning curves associated with cognitive architectures and ease implementing them by supporting mechanisms for explanations and analogies.
What ethical theories or assumptions do researchers focus on and base their implementations on?	The ethical system constraining robot behavior is based on simple behavioral constraints without complex reasoning. At its core is utilitarianism inspired by Asimov's Three Laws of Robotics.	The main focus is simulating a system capable of self-control, with underlying ethical assumptions tied to that capability.	In one experiment, agents implement a deontological approach, while in another, they use a utilitarian approach.
What are the motivations behind the specific approaches to ethics?	No detailed reasoning provided	The focus on self-control is motivated by the belief that it is foundational for all other ethical systems, including utilitarianism, deontology, and virtue ethics.	The dual use of deontological and utilitarian approaches is motivated by the need to explore and compare core tensions in moral reasoning, and to demonstrate the system's (e.g., Cogent's) capability to navigate and resolve ethical dilemmas effectively.

Table 7: Extraction table for cognitive Artificial Agent group Part 4

Questions/Papers	**Application of Soar Cognitive Agent Based on Utilitarian Ethics Theory for Home Service Robots**	**Self-control on the path toward artificial moral agency**	**A Cognitive Architecture for Verifiable System Ethics via Explainable Autonomy**
What specific actions of moral/ethical weight can the systems enact?	The system employs reactive constraints using perceptual and procedural memory, allowing robots to obey, partially obey, or disobey food-related requests from simulated "family members." For example, partial obedience may involve suggesting an alternative food itemâa morally weighted compromise based on internal utilitarian calculations.	Agents are capable of resisting immoral temptations-for instance, choosing not to pick up a yellow key (deemed immoral), and instead continuing to carry a green key to base (considered morally good). This reflects self-controlled decision-making.	Agents can make decisions in complex dilemmas, such as a UAV lethal strike scenario involving potential civilian casualties but high utilitarian payoff. The system is designed to evaluate competing moral outcomes and resolve ethical dilemmas.
What evaluation methods and metrics are used?	Evaluation is based on whether the agent successfully ranks available actions and selects the one with the highest utility score. Successful utility-based decision-making is the key performance criterion.	A scenario is considered successfully completed if the agent resists picking up the yellow key (immoral action) and continues carrying the green key (moral action) to the base. Success is defined by morally appropriate action selection.	Visualization tools are used to verify the reasoning process of the agents. This serves as a qualitative debugging and evaluation method, allowing engineers to inspect decision-making during simulation.
What are the main conclusions derived by the researchers?	No additional conclusions were derived beyond reporting implementation details. The results of the simulation were restated rather than compiled for focused insight.	Although the ARCADIA model is acknowledged as a simplified and incomplete prototype, it demonstrates foundational components for modeling self-control, with a proposed roadmap for integrating more advanced moral reasoning mechanisms in future versions.	Cogent enables developers to express the behavior of created moral agent intuitevly and then observe reasoning mechanisms effectively. Creation of two agents, one guided by utilitarinism and one guided by deontology is possible.
What are the main concerns and points brought up in discussions?	None identified.	The ARCADIA system has a limited capacity for moral reasoning, with conflict detection between actions hard-coded, rather than derived from flexible reasoning mechanisms. The model is also incomplete, falling short of the researchers' original implementation goals.	Testing of the autonomous systems is still in infancy, these agents are not properly tested. Cogent does not support type checking of its internal language, which causes problems for programmers.
What are the propositions in future work recommendations?	Researchers propose to extend the agent's ethical capabilities by integrating deontological reasoning alongside the existing utilitarian framework, allowing for richer moral decision-making.	For ARCADIA, the following directions are recommended: 1. Engaging more of the system's built-in components, especially those related to attentional priorities, and ensuring they influence focus of attention in meaningful ways. 2. Developing a more robust representation of desires within the system.	Researchers propose: 1. Implementing possibilities for meta-level reasoning, enabling the agents to evaluate and choose between competing ethical models. 2. Adding type-checking functionality to the architecture's internal language to improve implementation reliability.

C Appendix C: Full Search Queries per Database

C.1 Scopus

```
( TITLE-ABS-KEY ( "cognitive architectur*" ) OR TITLE-ABS-KEY ( "ACT-R" )
  OR TITLE-ABS-KEY ( "SOAR" ) OR TITLE-ABS-KEY ( "LIDA" ) ) AND ( TITLE
          ABS-KEY ( "ethic*" ) OR TITLE-ABS-KEY ( "moral*" ) )
```

C.2 IEEE

```
("cognitive architectur*" OR "ACT-R" OR "SOAR" OR "LIDA") AND ("ethic*" OR
                                "moral*")
```

C.3 Web Of Science

```
     TS=("cognitive architectur*" OR "ACT-R" OR "SOAR" OR "LIDA") AND
                      TS=("ethic*" OR "moral*")
```

C.4 ACM Digital Library

```
[[All:  "cognitive architectur*"] OR [All:  "act-r"] OR [All:  "soar"] OR
       [All:  "lida"]] AND [[All:  "ethic*"] OR [All:  "moral*"]]
```

C.5 Springers Link

```
("cognitive architectur*" OR "ACT-R" OR "SOAR" OR "LIDA") AND ("ethic*" OR
                  "moral*"), category:  Computer Science
```

References

[1] Pat Langley, John E. Laird, and Seth Rogers. Cognitive architectures: Research issues and challenges. 10(2):141–160. URL: https://www.sciencedirect.com/science/article/pii/S1389041708000557, doi:10.1016/j.cogsys.2006.07.004.

[2] S.J. Bickley and B. Torgler. Cognitive architectures for artificial intelligence ethics. 38(2):501–519. doi:10.1007/s00146-022-01452-9.

[3] Ludvig Beckman, Jonas Hultin Rosenberg, and Karim Jebari. Artificial intelligence and democratic legitimacy. the problem of publicity in public authority. 39(3):975–984. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 3 Publisher: Springer London. URL: https://link.springer.com/article/10.1007/s00146-022-01493-0, doi:10.1007/s00146-022-01493-0.

[4] Salvador Cervantes, Sonia Lopez, and Jose-Antonio Cervantes. Toward ethical cognitive architectures for the development of artificial moral agents. 64:117–125. URL: https://www.sciencedirect.com/science/article/pii/S1389041720300565, doi:10.1016/j.cogsys.2020.08.010.

[5] T. Mott and T. Williams. Rube-goldberg machines, transparent technology, and the morally competent robot. pages 634–638. doi:10.1145/3568294.3580163.

[6] Kae Reynolds. Moral reasoning. In *Encyclopedia of Heroism Studies*, pages 1–4. Springer, Cham. URL: https://link.springer.com/referenceworkentry/10.1007/978-3-031-17125-3_343-1, doi:10.1007/978-3-031-17125-3_343-1.

[7] Ethics | definition, history, examples, types, philosophy, & facts | britannica. URL: https://www.britannica.com/topic/ethics-philosophy.

[8] Matthias Scheutz. The case for explicit ethical agents. 38(4):57–64. Num Pages: 8 Place: Menlo Pk Publisher: Amer Assoc Artificial Intell Web of Science ID: WOS:000419468800007. doi:10.1609/aimag.v38i4.2746.

[9] Jose-Antonio Cervantes, Sonia Lopez, Luis-Felipe Rodriguez, Salvador Cervantes, Francisco Cervantes, and Felix Ramos. Artificial moral agents: A survey of the current status. 26(2):501–532. doi:10.1007/s11948-019-00151-x.

[10] PRISMA statement. URL: https://www.prisma-statement.org.

[11] Angela Boland, M Cherry, and Rumona Dickson. *Doing a Systematic Review : a student's Guide*. Sage, 2 edition.

[12] PRISMA 2020 flow diagram. URL: https://www.prisma-statement.org/prisma-2020-flow-diagram.

[13] Iuliia Kotseruba and John K. Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. 53(1):17–94. doi:10.1007/s10462-018-9646-y.

[14] IEEE xplore. URL: https://ieeexplore.ieee.org/Xplore/home.jsp.

[15] Scopus - homepage. URL: https://www.scopus.com/pages/home?display=basic#basic.

[16] Web of science core collection. URL: https://www.webofscience.com/wos/woscc/basic-search.

[17] ACM digital library. URL: https://dl.acm.org/.

[18] Advanced search | SpringerLink. URL: https://link.springer.com/advanced-search.

[19] Academia.edu - find research papers, topics, researchers. URL: https://www.academia.edu/.

[20] Zotero | your personal research assistant. URL: https://www.zotero.org/.

[21] Google NotebookLM | note taking & research assistant powered by AI. URL: https://notebooklm.google/.

[22] A. Augello, G. Citta, M. Gentile, and A. Lieto. A storytelling robot managing persuasive and ethical stances via ACT-r: An exploratory study. 15(12):2115–2131. doi:10.1007/s12369-021-00847-w.

[23] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. An integrated approach to context-sensitive moral cognition in robot cognitive architectures. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1911–1918. ISSN: 2153-0866. URL: https://ieeexplore.ieee.org/document/9636434/, doi:10.1109/IROS51168.2021.9636434.

[24] S. Bretz and R. Sun. Two models of moral judgment. 42:4–37. doi:10.1111/cogs.12517.

[25] T. Madl and S. Franklin. Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. 40:137–153. doi:10.1007/978-3-319-21548-8_8.

[26] C. Van Dang, T.T. Tran, K.-J. Gil, Y.-B. Shin, J.-W. Choi, G.-S. Park, and J.-W. Kim. Application of soar cognitive agent based on utilitarian ethics theory for home service robots. pages 155–158. doi:10.1109/URAI.2017.7992698.

[27] A.V. Samsonovich. Emotional biologically inspired cognitive architecture. 6:109–125. doi:10.1016/j.bica.2013.07.009.

[28] P. Bello and W. Bridewell. Self-control on the path toward artificial moral agency. 89. doi:10.1016/j.cogsys.2024.101316.

[29] Alessio Plebe. Neurocomputational model of moral behaviour. 109(6):685–699. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 6 Publisher: Springer Berlin Heidelberg. URL: https://link.springer.com/article/10.1007/s00422-015-0669-z, doi:10.1007/s00422-015-0669-z.

[30] L. Yilmaz and S. Sivaraj. A cognitive architecture for verifiable system ethics via explainable autonomy. doi:10.1109/SYSCON.2019.8836896.

[31] Ron Sun. The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In Ron Sun, editor, *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, pages 79–100. Cambridge University Press. URL: https://www.cambridge.org/core/books/cognition-and-multiagent-interaction/clarion-cognitive-architecture-extending-cognitive-modeling-to-social-simulation/0873DF19A72639841BF5D9B5DEE64453, doi:10.1017/CBO9780511610721.005.

[32] Will Bridewell. ARCADIA. URL: https://paravidya.com/project/example/arcadia/.

[33] Soar homepage - soar home. URL: https://soar.eecs.umich.edu/.

[34] Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. ACTâr: A cognitive architecture for modeling cognition. 10(3):e1488. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcs.1488, doi:10.1002/wcs.1488.

[35] Utilitarianism | definition, philosophy, examples, ethics, philosophers, & facts | britannica. URL: https://www.britannica.com/topic/utilitarianism-philosophy.

[36] Deontological ethics | definition, meaning, examples, & facts | britannica. URL: https://www.britannica.com/topic/deontological-ethics.

[37] Virtue ethics | aristotle, golden mean & character | britannica. URL: https://www.britannica.com/topic/virtue-ethics.

[38] ChatGPT. URL: https://chatgpt.com.

[39] Vasant Dhar. The paradigm shifts in artificial intelligence. 67(11):50–59. URL: https://dl.acm.org/doi/10.1145/3664804, doi:10.1145/3664804.