PhD

48

# NAI 2.0™:
# Prototype Deployment for Eldercare

## Academic & Scholarly Educational Research

A Hardware-Enforced Non-Agentic AI Governance Framework

Doctoral Thesis · NTU ARISE · 2026

**Edwin Koh Wui Kiat**

Founding Father · Non-Agentic AI 2.0™

ACRA T260229801 · SG020603109STW · 2026

□□ · □□ · □□ · □ · □□□□

NAI 2.0™ · STANDALONE REFERENCE · E-Book No. 48
Non-Agentic AI 2.0™ · kohedwin.ai · Academic & Scholarly Series

# NAI 2.0™:
# Prototype Deployment for Eldercare
## Academic and Scholarly Educational Research

A Hardware-Enforced Non-Agentic AI Governance Framework for Eldercare

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

**Edwin Koh Wui Kiat · Tiger**
Founding Father · Non-Agentic AI 2.0™ · kohedwin.ai
MBA, Maastricht School of Management · BSc ISE, University of Florida

**Submitted to NTU ARISE — Ageing Research Institute for Society and Education**
Nanyang Technological University · Singapore · 2026

謙虛 · 沉默 · 尊严 · 仁 · 止於至善

P-LIFE 1.00™ · Harm = Death · North = Save Life

ACRA T260229801 · Patent SG020603109STW · NLB R260219-005 

# Declaration of Originality

I hereby declare that this thesis is my own original work and that it has not been submitted for any other degree or qualification. All sources used have been acknowledged and cited in accordance with academic conventions.

The framework described in this thesis — Non-Agentic AI 2.0™ (NAI 2.0™) — and its associated components (ABC+2S+H™, 3ZEROS™ Protocol, Sacred Pause™, Tiger .1x Key™, Elder Dignity Score™, P-LIFE 1.00™) are the original intellectual property of the author, registered under ACRA entity T260229801 and patent IPOS SG020603109STW.

Where concepts, frameworks, or arguments from existing literature have informed the thesis, these have been carefully distinguished from the original contributions of this work.

Edwin Koh Wui Kiat · Tiger
Founding Father, Non-Agentic AI 2.0™
kohedwin.ai · ACRA T260229801
Singapore · 2026

# Abstract

This doctoral thesis develops and evaluates NAI 2.0™ (Non-Agentic AI 2.0™) as a hardware-enforced, non-agentic AI governance framework specifically designed for high-impact eldercare settings. The central problem addressed is the tendency of advisory AI systems to become functionally authoritative in care workflows, particularly in contexts involving vulnerable older adults where meaningful human sovereignty, dignity preservation, and accountability are indispensable requirements.

The thesis argues that policy-based AI governance — relying on configurable software guardrails, procedural oversight requirements, or institutional guidelines — is structurally insufficient to preserve human sovereignty in eldercare. Drawing on design science research, conceptual analysis, regulatory mapping, and feasibility evaluation, the study proposes a constitutional architecture in which governance constraints are embedded directly into the operational design of the AI system rather than imposed externally through policy.

The NAI 2.0™ framework is organized around four interconnected constitutional components: the 3ZEROS Sanctuary (privacy by physical architecture), the Sacred Pause™ (hardware-enforced deliberative delay), the Sovereign Brake (physical disconnection of AI from clinical action pathways), and the Tiger .1x Key™ (tripartite physical authentication for consequential authorization). Together, these components are designed to ensure that consequential AI outputs cannot move into clinical action without explicit, traceable human authorization.

The framework is evaluated against dimensions of conceptual coherence, architectural plausibility, operational feasibility, governance alignment, and dignity-preserving intent. It is mapped against relevant regulatory frameworks including AIHGle (Singapore), the EU AI Act, HIPAA, GDPR, and CCPA, demonstrating substantial alignment with emerging expectations for high-risk healthcare AI. The Elder Dignity Score™ is introduced as an exploratory instrument for assessing the degree to which AI deployment preserves elder psychological sovereignty.

The thesis concludes that NAI 2.0™ represents a coherent, governance-ready, and dignity-conscious prototype architecture for bounded eldercare AI. It makes original contributions in the areas of constitutional AI governance design, hardware-enforced human sovereignty, dignity-as-design-constraint, and care-specific AI regulatory alignment. Future work is identified in the areas of empirical clinical validation, large-scale feasibility testing, and international regulatory engagement.



**Keywords:** non-agentic AI, eldercare governance, hardware enforcement, human sovereignty, dignity preservation, Sacred Pause™, constitutional AI architecture, ABC+2S+H™, Singapore healthcare, HSA SaMD

# Contents

# Acknowledgements

This doctoral thesis represents twelve years of direct eldercare field observation, constitutional AI governance research, and patent development conducted across Singapore's hospitals, Active Ageing Centres, and community care settings. The work is dedicated to the 650+ elders who could not speak for themselves — whose silence is the reason this framework exists.

I acknowledge with deep gratitude the support and intellectual engagement of NTU ARISE — Ageing Research Institute for Society and Education — whose research infrastructure, clinical partnership framework, and ageing research mission align precisely with the constitutional governance principles developed in this work.

The constitutional values that govern this research — 謙虛 (humility), 沉默 (silence), 尊严 (dignity), 仁 (benevolence), 止於至善 (finish in the highest excellence) — were present in every design decision, every patent specification, and every line of this thesis. They are not rhetorical adornments. They are the operating system of the work.

Edwin Koh Wui Kiat

Singapore, 2026

# Chapter 1. Introduction

## 1.1 Background to the Study

Artificial intelligence is increasingly presented as a transformative force in healthcare. Across diagnostic support, triage, predictive analytics, workflow coordination, and patient monitoring, AI systems are being adopted with the promise of improving efficiency, consistency, and responsiveness. In many areas, these systems are no longer speculative. They are becoming part of routine care infrastructures, influencing how information is interpreted, how risks are prioritized, and how decisions are made. This expansion has been accompanied by strong claims about innovation, scalability, and the ability of AI to relieve pressure on overstretched health and care systems.

At the same time, the growth of AI in care environments has exposed a deeper governance problem. The central issue is not simply whether AI systems are accurate, useful, or cost-effective. It is also whether they are designed in ways that preserve meaningful human control, especially in settings involving vulnerable persons and high-impact decisions. Much contemporary AI discourse assumes that human oversight can be maintained if a clinician, caregiver, or operator remains somewhere in the process. Yet in practice, system outputs may acquire authority through workflow design, interface structure, institutional pressure, and habits of reliance long before formal autonomy is ever declared. The result is that advisory systems can become functionally decisive even while remaining nominally subordinate.

This concern is particularly acute in eldercare. Older adults may face dependency, frailty, cognitive fluctuation, sensory decline, chronic illness, social isolation, or complex combinations of medical and non-medical need. Care decisions in this context often involve not only safety and treatment, but privacy, dignity, routine, family involvement, and the preservation of personhood under conditions of increasing vulnerability. Unlike many acute medical settings, eldercare is not defined solely by episodic intervention. It is a domain of ongoing relationship, trust, attentiveness, and negotiated support. Technologies introduced into this domain therefore shape not only outcomes, but the character of care itself.

The appeal of AI in eldercare is understandable. Systems that assist with risk monitoring, event prediction, care coordination, documentation burden, and escalation logic may appear especially valuable in the face of demographic change, workforce shortages, and rising demand. However, these same systems may also intensify surveillance, depersonalize care, accelerate decision pathways without reflection, or displace human judgment in subtle but consequential ways. The introduction of AI into eldercare thus presents a tension between legitimate aspirations for support and the risk of allowing machine-driven logic to dominate contexts where dignity and human recognition are indispensable.

This thesis begins from the view that the governance challenge in eldercare is not solved by simply adding ethical principles to otherwise autonomy-seeking systems. Nor is it solved by assuming that healthcare regulation, in its current general form, is sufficient to preserve dignity-sensitive human control. What is needed is a more explicit and enforceable architecture of restraint: a way of designing AI systems so that they remain bounded, interruptible, review-dependent, and subject to materially meaningful human authority. This is the context in which

NAI 2.0™ is proposed.

NAI 2.0™ is developed in this thesis as a hardware-enforced, non-agentic AI governance framework for eldercare. It is intended not as an argument against digital support, but as an alternative to agentic or weakly governed AI models that risk converting decision support into practical authority. Its central premise is that eldercare AI should be designed constitutionally, with explicit structural controls that preserve human sovereignty, protect dignity, and prevent high-impact outputs from moving seamlessly into action without pause, review, and authorization.

This introductory chapter establishes the rationale for the study, defines the research problem, presents the aim and objectives, sets out the research questions, explains the significance and scope of the work, and outlines the structure of the thesis. It frames the dissertation as an interdisciplinary response to a pressing and insufficiently resolved question: how should AI be architected in eldercare so that it supports care without displacing accountable human authority?

## 1.2 Context and Rationale

The rationale for this study arises from the convergence of three developments.

First, AI systems are becoming increasingly embedded in care-relevant workflows. Even when systems do not make final decisions autonomously, they shape prioritization, attention, timing, and escalation. Their outputs may influence how professionals interpret risk, how institutions allocate resources, and how care staff respond to events. This means that AI can affect care trajectories even when described as "supportive" or "advisory."

Second, eldercare presents a uniquely sensitive deployment environment. Older adults may be more exposed to the effects of automation because they are often situated within dependency relationships and institutionally structured care pathways. They may not be well positioned to contest AI-influenced decisions, particularly where cognitive or communication impairments exist. Moreover, many ethically relevant aspects of eldercare cannot be reduced to optimization problems. The right course of action may depend on preference, biography, context, family dynamics, cultural expectations, or tolerance for risk. These are not variables that AI can reliably settle on its own.

Third, current governance discourse often remains too abstract. Frameworks for trustworthy AI commonly emphasize fairness, transparency, accountability, human

oversight, and privacy. These principles are important, but they are frequently stated at a level that does not determine actual system behavior. A system may formally comply with broad ethical language while still encouraging over-reliance, obscuring responsibility, or normalizing machine-led workflows. The practical question is therefore not whether such principles should exist, but how they can be translated into system design.

This thesis argues that the crucial missing element is bounded authority. Many current approaches assume that the challenge is to build more capable AI and then control the risks around it. By contrast, this study begins with the proposition that in eldercare, the first design question should be what the AI is not allowed to do. This reverses the usual trajectory of innovation discourse. Instead of asking how far autonomy can safely expand, it asks how constitutional constraints can preserve care as a fundamentally human-governed practice.

The rationale is strengthened by the fact that eldercare sits at the intersection of healthcare, social care, domestic life, and institutional governance. This makes it difficult to apply generic AI models without distortion. A narrowly clinical AI framework may overlook dignity and relational care. A consumer technology framework may underestimate the significance of vulnerability and accountability. A purely technical framework may fail to appreciate the power of workflow design in shaping care relationships. The present study therefore treats eldercare as a distinct site requiring a specifically tailored governance architecture.

In this context, NAI 2.0™ is developed as a proposed answer to a practical and normative need: a framework in which AI remains supportive, bounded, and reviewable, while high-impact transitions are protected through constitutional mechanisms such as 3ZEROS Sanctuary, Sacred Pause™, Sovereign Brake, and Tiger .1x Key™. These components are not presented as superficial features or branding devices, but as integral parts of a system intended to preserve meaningful human sovereignty over consequential care processes.

## 1.3 Problem Statement

Despite rapid growth in healthcare AI, there remains no sufficiently specified governance architecture for eldercare that ensures AI systems remain non-agentic, bounded, and human-sovereign in high-impact care contexts. Existing literature and practice tend to focus on predictive capability, decision support, workflow efficiency, or broad ethical principles, but they often fail to address a deeper and more operational problem: AI outputs can acquire practical authority without explicit authorization, particularly in environments where staff are time-constrained, users are vulnerable, and institutional processes favor speed and standardization.

This problem is not merely technical. It is sociotechnical, ethical, and regulatory. In eldercare, machine-generated outputs may influence escalation, monitoring intensity, prioritization, risk interpretation, or care pathway decisions that affect older adults' wellbeing, privacy, autonomy, and dignity. Yet current approaches to governance often

rely on procedural or rhetorical notions of "human oversight" that are insufficiently robust. A human may remain formally in the loop while lacking real opportunity, authority, or structural support to challenge, interrupt, or contextualize system outputs. Under such conditions, AI may become functionally agentic even if legally described as advisory.

The problem is compounded by the relative underdevelopment of eldercare-specific AI governance. While broader healthcare AI regulation is evolving, it often does not address the distinctive features of eldercare, including long-term dependency, fluctuating capacity, relational care, dignity sensitivity, quasi-domestic environments, and the ethical significance of non-coercive support. Similarly, ethical frameworks invoke autonomy, privacy, and respect, but often do not specify how these values should be embedded into the architecture of systems deployed in routine care settings.

As a result, there is a gap between principle and implementation. AI governance frameworks may declare the importance of accountability, human control, and safety, yet lack concrete design mechanisms that prevent authority drift. There is therefore a need for a governance model that does not merely recommend caution, but structurally enforces it.

This thesis addresses that gap by developing and evaluating NAI 2.0™ as a hardware-enforced, non-agentic AI governance architecture for eldercare. The problem it responds to may be stated directly as follows:

How can AI be designed for eldercare so that it provides meaningful support without acquiring unbounded practical authority over high-impact care processes?

## 1.4 Aim of the Study

The aim of this study is to develop and evaluate a hardware-enforced, non-agentic AI governance framework for eldercare that preserves meaningful human sovereignty, supports dignity-sensitive care, and aligns with emerging regulatory and ethical expectations for high-impact healthcare AI.

This aim reflects both a constructive and evaluative purpose. The thesis does not only critique current approaches; it proposes a new architectural model. At the same time, it does not present that model uncritically. It evaluates the framework conceptually, architecturally, ethically, and regulatorily to determine whether it offers a plausible and defensible alternative to more agentic forms of AI deployment in eldercare.

## 1.5 Research Objectives

To achieve this aim, the study pursues the following objectives:

1. To examine the current literature on AI in healthcare, eldercare technologies, human oversight, safety engineering, dignity, and AI governance in order to identify key conceptual and practical gaps.

2. To define the core concepts necessary for the study, including non-agentic AI, human sovereignty, constitutional architecture, hardware-enforced governance, and dignity preservation in eldercare.

3. To design NAI 2.0™ as a bounded governance architecture capable of constraining AI authority in eldercare through layered technical, procedural, and constitutional controls.

4. To specify the operational components of the framework, including 3ZEROS Sanctuary, Sacred Pause™, Sovereign Brake, Tiger .1x Key™, and the broader system layers that support bounded decision support.

5. To assess the framework against relevant regulatory and governance expectations, including healthcare AI lifecycle governance, Software as a Medical Device logic, high-risk AI oversight principles, and data governance frameworks.

6. To evaluate the framework's feasibility and interpretive strength in relation to operational workflow, architectural plausibility, human oversight, privacy, accountability, and dignity preservation.

7. To identify the limitations, residual risks, and future validation requirements associated with the framework as a doctoral-stage governance prototype.

## 1.6 Research Questions

The thesis is guided by one primary research question and a set of related sub-questions.

### Primary Research Question

How can a hardware-enforced, non-agentic AI governance architecture preserve meaningful human sovereignty and dignity in high-impact eldercare settings while remaining technically plausible and regulatorily aligned?

### Sub-Questions

8. What shortcomings exist in current AI, healthcare, and eldercare

governance literature regarding bounded authority and meaningful human control?

9. What conceptual foundations are necessary to define a non-agentic AI architecture suitable for eldercare?

10. How can governance principles such as pause, interruption, review, and explicit authorization be embedded into technical architecture rather than left at the level of policy alone?

11. To what extent does the proposed NAI 2.0™ framework align with regulatory, privacy, and ethical expectations relevant to healthcare and high-risk AI deployment?

12. How feasible is the framework as a workflow-sensitive model for real eldercare settings, and what tensions or limitations remain unresolved?

These questions structure the progression of the thesis from literature review and conceptual development to architecture design, governance mapping, and evaluative interpretation.

## 1.7 Central Thesis and Proposition

The central thesis advanced in this dissertation is that eldercare AI should be governed constitutionally rather than permissively. In other words, systems intended for high-impact care environments should be designed from the outset to preserve boundedness, interruptibility, explicit authorization, and traceable accountability, rather than assuming that expanding AI capability can later be moderated through procedural oversight alone.

The core proposition of the study is that a non-agentic, hardware-enforced governance framework can provide a more ethically serious, regulatorily legible, and operationally disciplined model for eldercare AI than architectures that permit seamless or weakly governed progression from machine output to care action.

This proposition rests on several supporting claims:

— that eldercare is a domain of distinctive vulnerability and dignity sensitivity;

— that nominal human oversight is often insufficient to prevent authority drift;

— that governance must be embedded into architecture, not only documented in policy;

— that high-impact AI outputs should be review-dependent and interruptible;

— and that bounded AI can still be useful without becoming authoritative.

The thesis does not claim that non-agentic design solves every problem associated with AI in care. Nor does it claim that NAI 2.0™ is already a fully validated deployment product. Rather, it claims that such a framework offers a stronger governance starting point for eldercare than prevailing autonomy-oriented assumptions.

## 1.8 Significance of the Study

The significance of this study lies in its contribution to scholarship, design practice, and governance thinking at a time when AI is entering increasingly sensitive healthcare and care environments.

### 1.8.1 Theoretical Significance

The study makes a theoretical contribution by advancing a clearer conceptual vocabulary for discussing AI authority in eldercare. Terms such as non-agentic AI, human sovereignty, and constitutional architecture are often absent or inconsistently used in current literature. By defining and integrating these concepts, the thesis provides a more precise framework for analyzing how AI should be bounded in dignity-sensitive care contexts.

It also contributes to debates on meaningful human control by shifting the focus from formal human presence to operationally real authority. In doing so, it extends discussions that are often more developed in military or algorithmic governance contexts into the specific environment of eldercare.

### 1.8.2 Practical and Design Significance

Practically, the study offers a concrete governance architecture rather than a purely critical account. NAI 2.0™ is significant because it attempts to show how governance can be embedded into the structure of an AI-enabled system through layered controls, role-based review, and escalation-sensitive friction. This is relevant to designers, healthcare innovators, digital health teams, and care providers seeking alternatives to more agentic system models.

### 1.8.3 Ethical Significance

Ethically, the study is significant because it centers dignity,

relational care, and human sovereignty in a domain where technological discourse often privileges efficiency, prediction, and scale. It argues that eldercare AI should not be judged solely by performance or throughput, but by whether it preserves the conditions under which older adults remain recognized as persons rather than managed objects of optimization.

### 1.8.4 Regulatory and Policy Significance

The study also contributes to governance and policy debates by exploring how a bounded AI architecture may better align with emerging risk-based regulatory expectations. Rather than treating regulation as a downstream compliance problem, the thesis models how governance logic can be built into the design itself. This makes the work relevant to discussions of high-risk AI, healthcare software governance, and privacy-sensitive deployment.

### 1.8.5 Significance for Eldercare

Most importantly, the study is significant because eldercare has been relatively under-theorized as a distinct AI governance context. Older adults occupy a position where health, dependence, domesticity, and institutional oversight converge. A framework designed specifically for this environment is therefore both timely and necessary.

## 1.9 Scope and Delimitations

This thesis is ambitious in conceptual reach, but it is also bounded in important ways. Its scope and delimitations should be stated clearly.

### 1.9.1 Scope

The study focuses on the design and evaluation of a governance architecture for AI in eldercare, particularly in contexts where system outputs may influence high-impact care processes. The thesis is concerned with:

— non-agentic AI design;

— human sovereignty and meaningful control;

— constitutional and hardware-enforced governance mechanisms;

— privacy, dignity, and accountability in eldercare;

— regulatory and ethical alignment of the framework;

— and feasibility-oriented interpretation of deployment logic.

The primary contribution is the development of NAI 2.0™ as a governance model, not the optimization of a single predictive algorithm.

### 1.9.2 Delimitations

Several delimitations apply.

First, the thesis does not present a large-scale clinical trial or a multi-site outcome study. Its claims are therefore strongest at the conceptual, architectural, and governance levels rather than at the level of demonstrated clinical superiority.

Second, the thesis does not claim that NAI 2.0™ is a certified medical device or legally approved product. Regulatory discussion is evaluative and mapping-based rather than determinative.

Third, the thesis does not attempt to solve all questions related to AI fairness, bias, or global health system variation, although these issues remain relevant. Its primary focus is bounded authority and governance in eldercare.

Fourth, the framework is designed for high-impact eldercare settings and may not apply in identical form to every health or social care context. Adaptation would be required for different infrastructures, jurisdictions, and care models.

Fifth, although the thesis introduces the Elder Dignity Score as an exploratory tool, it does not claim that dignity can be fully captured through quantitative measurement. The score is used cautiously and developmentally.

These delimitations are not weaknesses. They define the study's proper domain and ensure that its claims remain proportionate to the evidence presented.

## 1.10 Overview of Methodological Orientation

A full methodology is provided in Chapter 4, but a brief overview is useful here. The study adopts an interdisciplinary design science research orientation, supplemented by conceptual analysis, regulatory mapping, and feasibility-oriented evaluation. This

approach is appropriate because the thesis seeks to create and assess a governance artefact rather than test a single empirical variable.

The research proceeds by:

— identifying the governance gap in current literature and practice;

— synthesizing insights from AI, healthcare, eldercare, ethics, safety engineering, and regulation;

— developing the conceptual foundations of non-agentic and constitutional AI;

— designing the NAI 2.0™ architecture as a layered framework;

— evaluating the framework against ethical, technical, operational, and regulatory criteria;

— and interpreting its strengths and limitations as a thesis-stage governance prototype.

This methodological orientation reflects the nature of the problem. The challenge of eldercare AI is not solely technical, legal, or moral. It is a problem of sociotechnical design requiring multiple forms of reasoning at once.

## 1.11 Definition of Key Terms

For clarity, several key terms are introduced here in summary form. More detailed treatment appears in Chapter 3.

**Artificial Intelligence**

In this thesis, artificial intelligence refers broadly to computational systems capable of generating inferences, classifications, recommendations, or decision-support outputs from data in ways that may influence care processes.

**Non-Agentic AI**

Non-agentic AI refers to AI that does not possess or exercise independent executive authority over consequential care actions. Its outputs remain bounded, advisory, review-dependent, and subject to explicit human control.

**Human Sovereignty**

Human sovereignty refers to the preservation of materially effective human authority over high-impact decisions, especially the power to pause, review, reject, authorize, or redirect system-influenced care pathways.

**Constitutional Architecture**

Constitutional architecture refers to a system design approach in which core governance rules are embedded structurally into the operation of the system, shaping what it may and may not do.

**Hardware-Enforced Governance**

Hardware-enforced governance refers to the use of materially effective controls beyond ordinary software logic to ensure that certain transitions or actions cannot occur without explicit authorized human intervention.

**Dignity Preservation**

Dignity preservation refers to the maintenance of conditions consistent with respectful, non-coercive, privacy-conscious, person-centered care, including explanation, relational recognition, and protection against depersonalizing automation.

**Eldercare**

Eldercare is used here to refer broadly to organized support and care for older adults across institutional, residential, and assisted contexts where health, safety, daily living, and relational support intersect.

## 1.12 Structure of the Thesis

The thesis is organized into eight chapters, each contributing to the overall argument.

# Chapter 1: Introduction

This chapter introduces the background, rationale, problem statement, aim, objectives, research questions, significance, scope, and overall direction of the study.

# Chapter 2: Literature Review and Contextual Foundation

Chapter 2 examines the relevant literature on healthcare AI, eldercare technologies, automation, meaningful human control, dignity, privacy, safety engineering, and AI governance. It identifies the research gap addressed by the thesis.

# Chapter 3: Conceptual and Theoretical Framework

Chapter 3 defines the conceptual foundations of the thesis, including non-agentic AI, human sovereignty, constitutional architecture, and dignity-sensitive governance. It also introduces the core constitutional components of NAI 2.0™.

# Chapter 4: Methodology

Chapter 4 explains the study's methodological design, including its design science orientation, multi-phase development process, data sources, evaluation domains, exploratory Elder Dignity Score, and research ethics considerations.

# Chapter 5: NAI 2.0™ System Architecture and Deployment Model

Chapter 5 presents the architecture of the proposed framework in detail, including system boundaries, layered design, workflow logic, hardware-enforced controls, data handling, and deployment model.

## Chapter 6: Regulatory, Ethical, and Governance Alignment

Chapter 6 evaluates the framework against relevant regulatory, privacy, and governance expectations, including risk-based healthcare software logic, high-risk AI principles, and ethical governance concerns.

## Chapter 7: Evaluation, Feasibility, and Interpretation of Findings

Chapter 7 assesses the framework's overall strength as a governance prototype by examining conceptual coherence, architectural plausibility, workflow feasibility, regulatory readiness, dignity-related adequacy, and unresolved tensions.

## Chapter 8: Conclusion and Future Directions

The final chapter synthesizes the thesis's core contribution, restates its significance, identifies implications for eldercare AI governance, and sets out a future research and implementation agenda.

### 1.13 Chapter Summary

This introductory chapter has established the foundation for the thesis by identifying the central problem: the lack of a sufficiently bounded and eldercare-specific AI governance architecture capable of preserving meaningful human control in high-impact care settings. It has argued that the growing integration of AI into healthcare and eldercare creates not only opportunities for support, but also risks of authority drift, depersonalization, surveillance expansion, and dignity erosion if systems are not carefully constrained.

The chapter has presented the aim, objectives, and research questions guiding the study, and it has positioned the thesis as a constructive response to a real interdisciplinary gap. It has also clarified the significance, scope, and methodological orientation of the research, and introduced the central proposition that eldercare AI should be designed constitutionally rather than permissively.

The remainder of the thesis develops and evaluates this proposition through literature synthesis, conceptual clarification, methodological design, architectural specification, regulatory and ethical mapping, and feasibility-oriented interpretation. The next chapter begins that task by reviewing the literature and establishing the contextual foundation from which NAI 2.0™ emerges.

**Concluding Note to Chapter 1**

This chapter has introduced the thesis as an inquiry into how AI can be used in eldercare without allowing machine outputs to assume unbounded practical authority over vulnerable persons and dignity-sensitive care processes. It has argued that the central challenge is not merely technical performance, but governance: who or what remains in control when AI enters routine care pathways.

In response, the study proposes NAI 2.0™ as a hardware-enforced, non-agentic governance framework intended to preserve human sovereignty, support dignified care, and align with emerging expectations for high-risk healthcare AI. The significance of this work lies in its refusal to treat eldercare as a generic deployment environment or human oversight as a sufficient slogan. Instead, it approaches eldercare AI as a constitutional design problem requiring explicit architectural restraint.

With this foundation established, the thesis now turns to the existing body of scholarship in order to show more precisely what is known, what remains unresolved, and why a new governance model is necessary.

# Chapter 2. Literature Review and Contextual Foundation

## 2.1 Introduction

This chapter reviews the literature relevant to the development of NAI 2.0™ as a hardware-enforced, non-agentic AI governance framework for eldercare. Its purpose is to establish the scholarly context within which the thesis is positioned, identify the main conceptual and practical gaps in existing work, and justify the need for a new governance architecture specifically designed for high-impact eldercare settings.

The chapter proceeds from the premise that eldercare AI cannot be understood adequately through a single literature alone. The problem addressed by this thesis lies at the intersection of several domains: artificial intelligence in healthcare, digital technologies in eldercare, human oversight and meaningful control, biomedical ethics, safety engineering, data governance, and regulatory approaches to high-risk AI. Each of these literatures contributes something important, yet none on its own provides a sufficiently complete framework for constraining AI authority in eldercare while preserving dignity, accountability, and human sovereignty.

The review therefore has five main objectives. First, it examines how AI has been conceptualized and deployed in healthcare and care settings, with particular attention to decision support, risk prediction, monitoring, and workflow optimization. Second, it considers the growing use of digital systems in eldercare and the ethical, relational, and operational concerns these systems raise. Third, it reviews literature on automation, human oversight, and meaningful human control, identifying persistent weaknesses in current governance approaches. Fourth, it examines dignity, autonomy, vulnerability, and privacy as central normative concerns in eldercare rather than peripheral ethical additions. Fifth, it reviews the emerging regulatory and governance literature surrounding high-risk AI and healthcare software, showing where these discussions remain insufficiently connected to eldercare-specific design.

The central argument of this chapter is that while existing literature provides substantial insight into AI capability, healthcare regulation, and ethical principles, it remains underdeveloped in one critical respect: it does not adequately specify how AI systems in eldercare should be architecturally bounded so that machine output cannot silently become practical authority. Current scholarship often recommends human oversight, transparency, fairness, or accountability, but these are frequently expressed at the level of principle rather than embedded as operational constraints within system design. This gap is particularly serious in eldercare, where vulnerability, dependency, fluctuating capacity, dignity sensitivity, and asymmetries of institutional power make unbounded or weakly governed AI especially problematic.

For that reason, the chapter moves beyond descriptive review toward synthesis. It identifies the conceptual and practical conditions that motivate the thesis's proposed

contribution: a non-agentic, constitutional, hardware-reinforced framework capable of supporting eldercare AI without ceding high-impact authority to automated or quasi-automated processes. The chapter concludes by articulating the research gap that directly gives rise to NAI 2.0™.

## 2.2 Artificial Intelligence in Healthcare

The literature on AI in healthcare has grown rapidly over the past decade, driven by advances in machine learning, natural language processing, computer vision, and predictive analytics. Within this literature, AI is commonly presented as a tool capable of improving diagnostic support, risk prediction, triage, administrative efficiency, population health management, and clinical decision support. Much of the early enthusiasm for healthcare AI focused on the possibility that increasingly large data sets and increasingly sophisticated models could outperform conventional approaches in identifying patterns relevant to disease, deterioration, or treatment response.

A considerable portion of this scholarship is organized around technical performance. Studies frequently assess model accuracy, sensitivity, specificity, recall, precision, or area under the receiver operating characteristic curve. In many areas, this work has shown that AI can identify patterns in medical imaging, patient records, physiological signals, and workflow data with impressive technical capacity. Such literature has been important in demonstrating that AI systems can be clinically relevant rather than merely computationally novel.

However, a narrower performance-oriented focus also has limitations. Technical success in a retrospective or simulated dataset does not automatically translate into safe, appropriate, or acceptable real-world deployment. Healthcare is not a neutral information environment in which outputs flow seamlessly into action. It is a social, institutional, and ethical environment marked by uncertainty, competing responsibilities, legal constraints, contextual judgment, and unequal distributions of power. A model may perform well analytically while still producing unsafe or inappropriate effects when embedded in workflow, especially if its outputs are accepted too readily or integrated without adequate human scrutiny.

This concern has led a second stream of literature to focus on

clinical decision support systems rather than autonomous medical AI. Here, the emphasis shifts from pure model performance to how AI-generated outputs are interpreted, reviewed, and acted upon by healthcare professionals. This literature tends to recognize that AI in healthcare should often function as assistance rather than replacement. It also highlights problems such as alert fatigue, trust calibration, interface bias, deskilling, automation bias, and the tendency of users to over-rely on recommendations presented with undue authority.

Yet even this more governance-aware literature often stops short of fully specifying how human oversight should be made meaningful. Many systems are described as "human in the loop," but the reality of that loop may be thin. Humans may review outputs only after

the workflow has already been shaped by machine classification. They may face interface designs that frame acceptance as default. They may work under institutional conditions that reward speed over critical reflection. As a result, formal human presence does not necessarily amount to substantive human control.

This issue is especially relevant to the present thesis. Healthcare AI literature has generated substantial insight into capability, implementation, and risk, but it has often assumed that improved performance and retained human participation are sufficient safeguards. In practice, however, the deeper question concerns the distribution of authority between system and user. When an AI system influences escalation, prioritization, or recommended action in a high-impact setting, the distinction between "support" and "authority" may become blurred. The literature has identified this problem in fragments, but it has not consistently provided a governance architecture capable of resolving it.

Another important theme in healthcare AI literature is explainability and transparency. Considerable debate exists about whether AI systems should be interpretable, how much explanation is necessary for safe use, and whether post hoc explanations are adequate in clinical settings. These debates are relevant, but they are sometimes overburdened with responsibility. A system may be somewhat explainable and still unsafe if its workflow enables uncritical uptake. Conversely, a system may be difficult to explain in full technical detail yet still be responsibly deployable if strong review, interruption, and accountability mechanisms are built into its operational use. This suggests that explainability, while important, should not be treated as a substitute for structural governance.

The literature also increasingly acknowledges that AI systems must be evaluated throughout the lifecycle of design, deployment, monitoring, and revision. This is a significant development because it moves beyond the assumption that technical validity alone determines appropriateness. However, even lifecycle discussions often remain general. They identify desirable governance principles, such as monitoring, traceability, and human oversight, without sufficiently specifying how these are to be materially enforced in practice.

Taken together, the healthcare AI literature provides two important foundations for this thesis. First, it establishes that AI can meaningfully affect healthcare decisions and workflows. Second, it reveals that technical performance alone cannot settle questions of safe and ethical deployment. What remains less developed is a concrete model for bounded authority, particularly in contexts like eldercare where vulnerability, dignity, and contextual care relationships make the consequences of authority drift especially serious.

## 2.3 Digital Technologies in Eldercare

The literature on digital technologies in eldercare is diverse, spanning telecare, remote monitoring, assistive robotics, ambient intelligence, predictive deterioration systems, medication support technologies, and digital care coordination tools. Much of this work is motivated by demographic change, rising care demand, workforce shortages, and the

need to support older adults across domestic, residential, and institutional care environments. Technology is frequently presented as a response to capacity strain: a means of extending observation, improving coordination, reducing delay, and supporting independent living for longer.

A recurring theme in this literature is the promise of supportive monitoring. Sensors, wearable devices, smart home systems, and digital record platforms are often promoted as tools for identifying falls, behavioral changes, medication non-adherence, mobility decline, wandering risk, or other signals relevant to older adults' wellbeing. Such systems are often justified in terms of safety and efficiency. They may also be described as enabling early intervention or reduced burden on caregivers and health systems.

While these aims are understandable, eldercare literature also reveals a distinctive ethical and relational complexity not always present in broader healthcare technology debates. Older adults are not simply patients in episodic clinical interaction. They may be residents, long-term care recipients, family members, socially isolated persons, people living with dementia, or individuals whose care needs are deeply intertwined with daily routine, personal identity, and dependence on others. Technologies introduced into this environment do not merely improve or hinder performance; they can alter the lived conditions of care itself.

This is where the literature becomes especially important for the thesis. A significant body of work in eldercare warns that technologies designed for safety may also intensify surveillance, reduce privacy, undermine autonomy, or shift care relationships toward managerial control. Monitoring technologies may generate a sense of protection, but they may also produce feelings of constant observation or loss of domestic intimacy. Assistive systems may support independence, yet they may simultaneously narrow space for negotiated risk, personal preference, or human presence. The literature therefore makes clear that eldercare cannot be treated as a simple application domain for general healthcare AI. It involves a distinctive balance between protection and intrusion.

Another recurring concern in eldercare scholarship is the risk of

depersonalization. When technologies are introduced primarily to improve throughput, standardization, or managerial visibility, older persons may be treated less as individuals with histories and preferences and more as units of operational risk. This concern is particularly acute when digital systems classify older adults according to frailty, compliance, deterioration probability, behavioral risk, or resource intensity without sufficient attention to how those classifications affect human interaction and self-understanding. The literature thus suggests that eldercare technologies carry a dual burden: they must not only be effective, but must also preserve the relational and dignitary dimensions of care.

There is also a growing literature on AI and robotics in eldercare that examines companionship technologies, social robots, automated reminders, and intelligent support systems. Some of this work is optimistic, emphasizing reduced loneliness, increased engagement, and enhanced practical support. Other scholarship is more critical, questioning whether such systems create simulated care without genuine reciprocity, obscure institutional neglect, or normalize substituting machine interaction for human

relationship. These concerns are not identical to the governance problem addressed by this thesis, but they share an important premise: eldercare technologies should not be evaluated solely in terms of functional performance.

A further strand of the literature focuses on implementation realities. It shows that older adults, family members, and caregivers often respond to digital systems in complex and sometimes ambivalent ways. Trust depends not only on capability, but also on usability, explanation, intrusiveness, and whether the system appears to support rather than displace human care. Staff adoption may also be uneven, particularly where technologies increase documentation burden, generate excessive alerts, or fail to fit existing routines. These observations are directly relevant to the feasibility dimension of the present thesis. A governance framework that introduces friction may be ethically justified, but it must also be operationally legible to those using it.

What emerges from the eldercare technology literature is not a rejection of innovation, but a warning against simplistic assumptions. Older adults often occupy positions of heightened vulnerability, yet they are also persons with preferences, histories, rights, and dignity claims that cannot be subordinated entirely to institutional efficiency. Existing literature identifies many of these tensions, but it has not yet fully resolved how AI systems can be designed so that support functions do not become covert forms of authority. This unresolved issue becomes even more pressing when AI outputs influence escalation, risk review, monitoring intensity, or care prioritization. In such contexts, the question is no longer whether technology is present, but how its power is structured.

## 2.4 Decision Support, Automation, and the Problem of Authority Drift

A key body of literature relevant to this thesis concerns decision support, automation, and the phenomenon often described as automation bias or over-reliance on system outputs. In healthcare, aviation, industrial safety, and other complex domains, researchers have long observed that human operators do not simply use automated systems as neutral tools. Instead, automation shapes attention, expectations, timing, and perceived legitimacy. Even where humans remain formally responsible, the structure of a system can make machine outputs difficult to challenge in practice.

In clinical decision support literature, this problem appears in several forms. One is the tendency for users to accept system recommendations because they are embedded in authoritative interfaces or institutional workflows. Another is the risk that frequent exposure to alerts conditions staff either to trust the system too readily or to ignore it habitually, producing a damaging oscillation between over-trust and fatigue. A further issue is that recommendations presented as operationally convenient may acquire practical force even when their evidentiary basis is incomplete or their relevance to the individual case is uncertain.

The literature on automation bias is especially important because it challenges simplistic accounts of "keeping a human in the loop." Human presence alone does not guarantee

human judgment. A person may technically click approval while substantively following a pathway heavily shaped by automation. If organizational norms, interface design, and time pressure all encourage assent, the human role may become one of symbolic endorsement rather than reflective decision-making.

This insight is central to the thesis. Much existing governance discourse assumes that responsibility can be preserved if a human remains somewhere in the process. Yet literature across safety-critical domains suggests that the real issue is not location alone, but

timing, authority, and interruptibility. To be meaningful, oversight must occur before consequential progression, it must be backed by actual power to stop or alter workflow, and it must be supported by system design rather than contradicted by it.

Related literature on decision support also raises the issue of

default pathways. Systems often produce outputs that are nominally optional but operationally privileged. For example, a recommendation may appear at a point in workflow where deviation requires more effort than acceptance. Or the user may be asked to justify rejection but not acceptance. Such design choices subtly redistribute authority in favor of the machine. They do not amount to autonomy in a formal sense, yet they may have similar practical effects.

Another concept relevant here is function creep. Systems initially introduced as support tools may gradually acquire wider roles as institutions become accustomed to relying on them. A predictive score may begin as informational but later drive prioritization, escalation, or resource allocation more directly. In highly pressured environments, this drift can occur without clear public deliberation. The literature suggests that governance structures must therefore anticipate not only initial misuse, but the gradual expansion of machine authority through routine practice.

The problem is compounded in eldercare because many consequential judgments are inherently contextual. A signal of agitation, deterioration, or non-compliance may have very different meaning depending on communication barriers, pain, confusion, culture, family presence, or the older person's prior wishes. Automated recommendations in such cases can never be wholly self-sufficient. Yet if workflow design treats them as near-decisive, contextual judgment is displaced.

The literature thus supports an important conclusion: if the aim is to preserve meaningful human control, governance cannot rely solely on declarations of oversight. It must instead shape the architecture of decision support itself. This means designing systems in which AI outputs are bounded, impact-sensitive, interruptible, and traceable. Current literature recognizes parts of this need, but often leaves the operational solution underdeveloped. That unresolved space is one of the key motivations for the framework advanced in this thesis.

## 2.5 Human Oversight and Meaningful Human Control

The concept of human oversight has become central to AI governance debates, particularly in healthcare and high-risk domains. Regulatory and ethical frameworks commonly require that humans remain involved in reviewing or supervising AI outputs, especially where those outputs may affect health, rights, or access to essential services. The underlying intuition is understandable: if AI systems can be fallible, opaque, or context-insensitive, then human oversight provides a safeguard against inappropriate automation.

However, a major challenge in the literature is that human oversight is often underspecified. It may refer to any human involvement at all, from passive awareness of a system's existence to active intervention in real time. This ambiguity has led scholars increasingly to distinguish between formal oversight and meaningful human control. The former is satisfied when a human is technically present in the process. The latter requires that the human has sufficient information, timing, authority, and practical ability to shape or halt the relevant outcome.

Literature on meaningful human control is especially developed in discussions of autonomous weapons, automated decision systems, and algorithmic governance, but its underlying principles are transferable to healthcare. Several themes recur. First, meaningful control requires that the human understands the nature and significance of the decision at stake. Second, it requires that the human intervention occurs at a stage where it can still affect the outcome. Third, it requires that the human is not overwhelmed by system complexity, interface design, or institutional pressure. Fourth, it requires that the human's authority is genuine rather than nominal.

These insights are highly relevant to eldercare AI. A system may claim to support care while still undermining meaningful control if its outputs are difficult to question, if review occurs too late, or if users are subtly steered toward default acceptance. The literature therefore suggests that human control must be examined not only at the level of legal accountability, but at the level of interface design, workflow sequencing, institutional norms, and escalation logic.

Another important theme in this literature is the distinction between

oversight as monitoring and oversight as governance. Monitoring implies watching or checking what a system does. Governance goes further: it structures what the system is allowed to do in the first place. This distinction is crucial for the thesis. Many existing systems allow humans to monitor outputs after generation but do not constrain how those outputs enter workflow. A governance-oriented approach instead asks how to ensure that some outputs cannot become consequential without explicit review, pause, or authorization.

The literature also notes that meaningful human control cannot be reduced to explanation alone. Better explanations may help users, but a well-explained system can still exert excessive authority if its outputs drive action too quickly or if rejection is institutionally

costly. This reinforces the thesis's broader argument that explanation is valuable but insufficient. Human control must be supported by structural conditions.

There is, however, a limitation in the existing literature. While the need for meaningful oversight is widely recognized, many discussions remain abstract. They describe desired qualities of oversight but provide less detail about how these qualities should be embedded into actual technical and operational systems. This gap becomes even more noticeable in healthcare and eldercare, where governance must contend with urgency, workload, privacy, and sensitivity to individual circumstances.

Thus, the literature on human oversight provides an essential normative and analytical foundation for the present study, but it does not fully solve the architectural problem. It clarifies why human control matters and what qualities it should possess, yet it leaves open the question of how such control can be made non-bypassable, role-specific, and operationally real in high-impact eldercare AI. NAI 2.0™ is proposed as one answer to that unresolved question.

## 2.6 Dignity, Vulnerability, and Relational Care in Eldercare

Any serious examination of eldercare must engage with the literature on dignity, vulnerability, and relational care. These themes are not secondary ethical concerns added after technical design. They shape the very conditions under which eldercare technologies can be judged appropriate or inappropriate.

The concept of dignity has been interpreted in multiple ways in healthcare and bioethics literature. Some accounts emphasize inherent human worth. Others focus on autonomy, recognition, bodily integrity, privacy, respect, or freedom from humiliation and degradation. In eldercare, dignity often appears as both a moral principle and an experiential condition. It concerns not only whether a person is treated according to formal rights, but whether care practices recognize them as a person rather than as a burden, risk object, or passive recipient of institutional management.

This literature matters for AI governance because technological systems can affect dignity indirectly as well as directly. A system need not insult or coerce a person in order to undermine dignity. It may do so by narrowing decision space, intensifying surveillance, depersonalizing interaction, or shifting attention away from the person toward the system's classifications. In care settings, small workflow changes may carry significant dignitary implications. For example, a system that escalates decisions without adequate explanation may leave the older person feeling managed rather than heard. A monitoring regime that is justified as safety-enhancing may also alter the sense of privacy and home. A predictive label may shape how staff interpret behavior, reducing openness to the person's own account.

The literature on vulnerability deepens this concern. Older adults may experience physical frailty, chronic illness, cognitive impairment, loneliness, dependency, grief, or fluctuating

capacity. Vulnerability here should not be understood as a fixed essence of old age, but as a situational condition shaped by social, institutional, and bodily factors. This is important because AI systems deployed in eldercare do not operate over abstract data subjects. They affect persons whose ability to resist, contest, or interpret system-driven decisions may be limited by circumstance.

Relational care literature further complicates the picture. Care is not merely the delivery of efficient interventions. It is also a relationship involving trust, attentiveness, explanation, continuity, responsiveness, and often family involvement. Technologies that improve surveillance or speed may still weaken care if they reduce time for human interaction, displace attention from the person to the dashboard, or encourage staff to privilege system-generated categories over lived knowledge. This literature does not imply that technology is inherently anti-relational, but it insists that technologies be assessed in relation to the human relationships they shape.

There is also important work on autonomy in eldercare, particularly on the tension between safety and self-determination. Older adults may wish to accept risks that institutions seek to minimize. They may value privacy, routine, or independence in ways that do not align neatly with managerial notions of optimal safety. AI systems that detect risk or recommend intervention may therefore encounter situations where the ethically appropriate response is not simple maximization of prevention, but negotiated and contextual judgment.

This insight is highly relevant to the thesis's emphasis on Sacred Pause™ and human sovereignty. In eldercare, the right response is often not whatever appears most efficient or technically rational. It may depend on the older person's preferences, family context, legal status, communication needs, or broader care goals. Dignity-preserving systems must therefore leave room for deliberation, not merely generate optimized next steps.

At the same time, dignity literature also warns against treating dignity as too vague to guide design. While dignity cannot be reduced fully to a metric, it can still inform system requirements. Privacy proportionality, non-coercive review, explanation, preserved human presence, and refusal pathways are all design-relevant expressions of dignity. The thesis builds on this line of thought by treating dignity as a system-level design constraint, not merely an ethical aspiration.

The literature thus provides a strong basis for insisting that eldercare AI be designed differently from purely efficiency-oriented systems. Yet it often stops at ethical critique rather than architectural specification. It tells us why dignity matters, but less often how governance structures can make dignity more likely in practice. That gap becomes a central driver of the present study.

## 2.7 Privacy, Datafication, and Surveillance Concerns

Privacy and data governance occupy a central place in contemporary literature on digital health and care technologies. In eldercare, these issues are particularly sensitive because

technologies often collect intimate, continuous, and behaviorally revealing forms of data. Unlike episodic medical information captured during clinical encounters, eldercare technologies may monitor movement, routines, sleep, speech, medication patterns, location, emotional state, or interactions within quasi-domestic environments. As a result, privacy concerns in eldercare are not limited to data breach or unauthorized access; they also involve the broader question of how much of a person's life becomes available for observation and institutional interpretation.

A key theme in the literature is datafication. This refers to the transformation of lived experience into data points that can be stored, analyzed, categorized, and acted upon. In healthcare, datafication is often treated as a source of insight and improved personalization. In eldercare, however, the literature raises sharper questions about what is lost when complex human situations are rendered as risk markers, behavioral anomalies, or system events. Datafication can improve visibility, but it can also narrow the meanings attributed to a person's conduct by privileging measurable indicators over narrative, context, and relationship.

Another important theme is surveillance creep. Systems introduced for limited protective purposes may gradually become normalized as routine observation mechanisms. For example, technologies initially justified by fall prevention or nighttime safety may become part of a broader culture of constant monitoring. The literature suggests that such expansion is not always the result of explicit policy change. It may occur incrementally, supported by managerial convenience, risk aversion, or the assumption that more data always means better care.

This has important implications for AI governance. AI systems do not merely process data; they create incentives for collecting data. More granular input may appear desirable for better prediction, but expanded collection also increases the risk of intrusion, misuse, and dignitary harm. This is particularly significant in eldercare, where privacy is often closely tied to personhood, modesty, and control over one's daily environment. Continuous monitoring may be tolerated for safety reasons, but it may still alter the experience of living with care.

The literature also examines access control, consent, purpose limitation, and proportionality. These concepts are widely recognized in legal and ethical governance, but eldercare complicates them. Some older adults may have diminished capacity. Others may rely on family members or institutional representatives in decisions about technology use. Consent may therefore be necessary but insufficient as a sole legitimating mechanism. Scholars have argued that privacy governance in care settings must also consider fairness, dependency, power imbalance, and the cumulative burden of observation.

There is also a growing critique of opaque secondary use. Data gathered in care settings may be repurposed for analytics, training, procurement evaluation, or administrative oversight in ways that are not obvious to older adults or even frontline staff. This issue is especially relevant where AI is involved, since system improvement often depends on large-scale data use. The literature thus emphasizes the need for strong purpose limitation and transparent governance around how care data are handled.

From the perspective of the present thesis, privacy literature supports two major conclusions. First, eldercare AI should adopt a

minimum-necessary data logic rather than a maximalist capture model. Second, privacy protection cannot be treated as a separate legal compliance issue disconnected from architecture. The structure of data intake, processing, access, retention, and auditability is itself a governance matter. Existing literature identifies this need clearly, but again often at the level of principle. It leaves open how privacy-conscious restraint should be integrated into the design of bounded AI systems for eldercare. NAI 2.0™ addresses this by treating privacy as both a technical and dignitary requirement within the system architecture.

## 2.8 Safety Engineering, High-Reliability Thinking, and Interruptibility

The literature on safety engineering and high-reliability systems offers another important lens for the thesis. Although much of this work originates outside eldercare, including aviation, nuclear systems, industrial control, and broader patient safety research, its principles are highly relevant to AI deployment in contexts where errors can have serious consequences.

One key insight from safety engineering is that harm often arises not from isolated technical failure alone, but from the interaction between humans, tools, procedures, interfaces, and organizational pressures. This systems-oriented perspective is valuable because it shifts attention away from individual blame and toward the architecture of sociotechnical risk. In the context of AI, this means that a model's error rate is only one part of the safety picture. Other factors include timing of alerts, default settings, escalation pathways, workload conditions, bypass options, maintenance failures, and the visibility of uncertainty.

A further theme in safety literature is the importance of defense in depth. Safe systems often rely on multiple layers of protection rather than a single control. If one safeguard fails, another should remain available to interrupt harmful progression. This principle directly informs the present thesis. A governance model that depends solely on user vigilance or solely on software prompts is fragile. By contrast, a layered model incorporating bounded outputs, pause conditions, override mechanisms, logging, and stronger authorization logic is more consistent with established safety thinking.

The concept of interruptibility is also highly relevant. In safety-critical systems, it is often essential that a process can be halted when uncertainty, anomaly, or elevated risk arises. Interruptibility matters not only in emergencies but also in ambiguous situations where continuation would be imprudent without review. This supports the thesis's argument that certain AI outputs in eldercare should trigger meaningful pause rather than seamless continuation. If an output could materially affect bodily welfare, privacy, autonomy, or dignity, then the system should not be designed for automatic momentum.

Patient safety literature in healthcare similarly emphasizes escalation pathways, incident learning, auditability, and the management of near misses. These concepts underscore

that safe deployment requires visibility into what happened, why it happened, and where intervention was possible or absent. AI systems lacking robust logging and clear transition points make such analysis difficult. This is especially problematic where machine recommendations influence care but responsibility becomes diffuse.

Another useful contribution from safety engineering is the distinction between work as imagined and work as done. Formal procedures may describe careful review and orderly escalation, yet real-world practice under pressure may diverge substantially. This insight is highly relevant to eldercare AI. Even well-intentioned systems may be used differently from how designers expect, especially in resource-constrained environments. Therefore, a credible governance architecture must anticipate the possibility of drift, shortcutting, normalization of override, or ritualized approval.

The literature also cautions that excessive safety friction can itself create risk if it leads to burden, fatigue, or bypass behavior. Safety controls must be proportionate and intelligently targeted. This point is essential for the present thesis because a pause-based governance model could become counterproductive if implemented indiscriminately. The solution is not to reject friction altogether, but to reserve stronger constraints for outputs whose impact justifies them.

One area where the safety literature is less developed in relation to AI is the use of hardware-enforced governance in care settings. While physical interlocks, keyed access, or device-level constraints are well established in other safety domains, their application to AI-mediated decision workflows in eldercare is not yet deeply explored. This is one point at which the thesis seeks to extend existing thinking. The argument is that software prompts alone may be too weak in high-pressure care environments, and that stronger forms of materially enforced authorization may sometimes be necessary to preserve meaningful human sovereignty.

Overall, safety engineering literature reinforces the thesis's claim that trustworthy AI in eldercare requires more than accurate models and good intentions. It requires layered controls, interruptibility, fallback processes, and a serious understanding of how risk emerges in sociotechnical systems. What remains insufficiently specified in current literature is how to adapt these insights into a coherent eldercare AI governance framework. That is precisely the role NAI 2.0™ is intended to play.

## 2.9 Regulatory and Governance Literature on High-Risk AI

Recent years have seen rapid growth in regulatory and governance literature relating to AI, especially in sectors where system outputs may affect health, safety, rights, or access to essential services. This literature includes legal scholarship, policy analysis, regulatory guidance, and standards-oriented discussions surrounding healthcare AI, software as a medical device, algorithmic accountability, and high-risk AI governance. It forms an essential part of the context for this thesis because NAI 2.0™ is explicitly positioned not merely as a technical concept, but as a governance-oriented architecture.

A common feature of this literature is the move toward risk-based regulation. Rather than treating all AI systems alike, emerging governance models seek to differentiate between low-risk and high-risk applications based on domain, impact, potential harm, and the degree of influence over consequential decisions. Healthcare and care-related uses frequently fall toward the more heavily governed end of this spectrum because they can materially affect health status, privacy, autonomy, and welfare.

The literature on software as a medical device is especially relevant where AI systems provide information used for clinical or care decisions. Here, key themes include intended use, significance of information provided, clinical context, human oversight, documentation, change management, validation, and post-market monitoring. This work is important because it emphasizes that governance is not simply about whether software exists, but about what role it plays in the decision pathway and how that role is controlled.

Parallel to this, broader AI governance literature has increasingly stressed human oversight, traceability, transparency, risk management, and accountability. These principles appear repeatedly across policy frameworks and governance proposals. They are clearly valuable, but the literature often remains programmatic. It identifies what should be present without fully showing how such expectations are operationalized in system architecture, especially where high-impact workflow decisions are involved.

The regulatory literature also increasingly recognizes the special relevance of vulnerable persons. Systems affecting children, patients, disabled persons, or older adults may attract stronger ethical and legal scrutiny because the stakes of misuse, overreach, or opacity are higher. This is an important development, but eldercare-specific governance remains comparatively underdeveloped. Older adults are often subsumed under broader categories such as patients, consumers, or care recipients, which can obscure the distinctive interplay of long-term dependency, dignity, domesticity, and fluctuating capacity that characterizes many eldercare settings.

Data protection literature also contributes important governance principles, especially regarding minimization, purpose limitation, lawful processing, transparency, access control, retention, and accountability. As noted earlier, these principles are highly relevant to eldercare AI. However, much of the literature treats them in isolation from workflow governance. It focuses on what data may be processed, but less on how data-driven outputs are prevented from exerting undue authority in practice.

Another important issue in governance scholarship is auditability. Scholars increasingly argue that high-risk AI systems must produce records sufficient to support investigation, explanation, contestability, and quality control. This supports the thesis's emphasis on logging and traceable authorization. Yet auditability, too, has limits if it remains purely retrospective. Records help explain what happened after the fact, but they do not themselves prevent harm in the moment. This is why the present thesis argues for combining traceability with interruptibility and bounded activation.

A further strand of literature examines ethics washing or symbolic compliance. This critique warns that organizations may adopt the language of trustworthy or human-centered AI

while preserving practices that continue to centralize power, weaken accountability, or obscure meaningful contestability. This critique is directly relevant here. If eldercare AI is to be ethically serious, its governance claims must be structurally supported rather than rhetorically asserted.

The most important conclusion from the regulatory and governance literature is therefore twofold. First, there is increasing consensus about the kinds of controls high-risk AI should involve. Second, there is still insufficient detail about how these controls should be instantiated in eldercare-specific architectures. Governance literature tells us what values and procedural features matter, but it often leaves unresolved how to design a system in which those values are materially upheld. The present thesis enters precisely at that point of unresolved translation.

## 2.10 Gaps in the Existing Literature

The preceding sections reveal that current scholarship is rich but fragmented. Healthcare AI literature addresses technical performance and clinical decision support. Eldercare literature emphasizes vulnerability, relational care, and dignity. Safety engineering highlights layered protection and interruption. Regulatory scholarship stresses risk management, traceability, and human oversight. Yet these strands are rarely integrated into a single eldercare-specific governance framework.

Several key gaps emerge.

### 2.10.1 Lack of Eldercare-Specific AI Governance Architecture

The first and most obvious gap is the absence of a clearly articulated

eldercare-specific AI governance architecture. Existing literature discusses ethics, regulation, and technology adoption in eldercare, but often does so descriptively or normatively rather than architecturally. There is relatively little work specifying how system design itself should embody the constraints required by dignity-sensitive and high-impact care environments.

### 2.10.2 Under-Specification of Meaningful Human Control

A second gap concerns the persistent under-specification of meaningful human control. Human oversight is frequently recommended, but the literature often fails to distinguish between nominal involvement and materially effective authority. There is insufficient attention to how pause, timing, override, and explicit authorization can be structurally enforced within workflow.

### 2.10.3 Insufficient Attention to Authority Drift

A third gap is the limited treatment of authority drift, meaning the process by which AI recommendations become practically authoritative without formal delegation of power. Existing work discusses automation bias and over-reliance, but there is less systematic attention to how architecture can prevent outputs from silently shaping decisions beyond their intended scope.

### 2.10.4 Weak Integration of Dignity Into Technical Design

A fourth gap lies in the treatment of dignity. Although dignity is widely recognized as important in eldercare, it is often left at the level of ethical commentary. There is relatively little work translating dignity into design-relevant conditions such as pause space, explanation, privacy proportionality, human presence, non-coercive workflow, and interruptibility.

### 2.10.5 Limited Use of Safety Engineering Logic in Care AI Governance

A fifth gap concerns the relatively weak translation of safety engineering principles into eldercare AI governance. Healthcare AI literature increasingly acknowledges risk, but it does not always make full use of concepts such as defense in depth, work-as-done analysis, layered controls, or materially enforced interruptibility.

### 2.10.6 Lack of Discussion of Hardware-Enforced Governance

A sixth and particularly distinctive gap is the limited exploration of

hardware-enforced governance. While physical and infrastructural safety controls are common in other high-reliability domains, there is comparatively little scholarship examining whether high-impact AI workflows in care settings should sometimes be constrained by stronger-than-software authorization mechanisms.

### 2.10.7 Incomplete Synthesis Between Regulation and Architecture

Finally, there is a gap between regulatory principle and architectural implementation. Governance literature increasingly identifies desirable properties such as transparency, traceability, and human oversight, but often stops short of describing how these should be built into the structure of a system intended for use with vulnerable older adults.

These gaps justify the central contribution of this thesis. NAI 2.0™ is proposed not as a generic ethical AI model, but as an attempt to synthesize these fragmented literatures into a concrete governance architecture for eldercare.

## 2.11 Positioning the Present Study

Against this background, the present study is positioned as an interdisciplinary intervention into the literature rather than a narrow technical contribution. It draws from healthcare AI, eldercare ethics, safety engineering, and regulatory governance, but it is not reducible to any one of these fields. Its distinctive contribution lies in proposing that eldercare AI should be treated as a constitutional problem of bounded authority rather than primarily as a problem of capability optimization.

This positioning has several implications.

First, the thesis does not ask whether AI can be useful in eldercare. Existing literature already makes clear that digital systems can assist monitoring, coordination, and decision support. The more urgent question is what governance architecture is required if such systems are to be used without undermining dignity, privacy, and human sovereignty.

Second, the thesis does not reject AI on the grounds that it is inherently incompatible with care. Rather, it argues that the acceptability of AI depends on how authority is distributed, how interruption is enabled, how data are bounded, and how high-impact outputs are governed. This allows the study to move beyond both uncritical technological optimism and blanket technological pessimism.

Third, the study treats eldercare as a distinct governance context. It resists the assumption that frameworks designed for generic healthcare AI can simply be transferred unchanged into environments marked by long-term dependency, relational care, fluctuating capacity, and heightened dignity concerns. This eldercare specificity is one of the main ways the thesis seeks to advance the literature.

Fourth, the study positions hardware-enforced and constitutionally bounded design as a serious alternative to the common assumption that software prompts and procedural oversight are sufficient. In doing so, it extends existing governance discussions by arguing that some forms of control need to be built more deeply into the operating conditions of the system.

Finally, the study adopts a deliberately non-maximalist stance toward AI authority. Rather than treating greater autonomy as the natural endpoint of progress, it argues that in eldercare, value may lie in disciplined limitation. This is a significant departure from much innovation discourse and one of the key normative commitments of the thesis.

In this sense, the literature review does more than summarize prior work. It establishes the intellectual terrain on which the thesis stands and shows why a framework such as NAI 2.0™ is needed.

**Concluding Note to Chapter 2**

This chapter has reviewed the major bodies of literature relevant to the thesis and shown that the problem addressed by NAI 2.0™ arises from a genuine and unresolved scholarly gap. Research on AI in healthcare demonstrates the growing capability and significance of intelligent systems, but also reveals the limits of performance-based evaluation and the inadequacy of purely nominal human oversight. Literature on eldercare technologies highlights the distinct ethical and relational sensitivity of care environments involving older adults, especially in relation to dignity, privacy, vulnerability, and depersonalization. Work on automation, meaningful human control, safety engineering, and high-risk AI governance further reinforces the need for bounded, interruptible, and traceable system design.

Across these literatures, a consistent pattern emerges. Principles such as oversight, accountability, privacy, and dignity are widely endorsed, yet they are often left insufficiently operationalized. What remains underdeveloped is a concrete eldercare-specific architecture that translates these principles into enforceable constraints on AI authority. This gap is especially significant because eldercare settings are precisely those in which authority drift, surveillance expansion, and dignity erosion can occur quietly through routine workflow rather than through dramatic technical failure.

The review therefore justifies the central direction of the thesis. A new model is needed---one that does not simply call for responsible AI in general terms, but specifies how AI in eldercare can be made non-agentic, bounded, interruptible, and sovereignty-preserving in practice. This is the contribution that the following chapter begins to develop by setting out the conceptual foundation of NAI 2.0™ as a constitutional framework for sovereign eldercare governance.

# Chapter 3. Conceptual and Theoretical Framework

## 3.1 Introduction

This chapter establishes the conceptual and theoretical basis of the thesis by defining the key ideas that underpin NAI 2.0™ and explaining how they are organized into a coherent governance architecture for eldercare. The central argument of the thesis is that, in high-impact care environments involving older adults, AI should remain structurally bounded rather than operationally autonomous. In this context, the purpose of system design is not to maximize independent machine action, but to ensure that human authority, contextual judgment, and dignity-preserving care remain primary.

The need for conceptual clarity is especially important in a study such as this one because terms such as non-agentic AI, sovereignty, constitutional architecture, and dignity are often used in broad or inconsistent ways across technical, ethical, and regulatory literature. If these terms are left vague, the framework risks appearing rhetorical rather than analytical. Accordingly, this chapter defines each concept in a thesis-specific manner and then locates the proposed framework within relevant bodies of theory, including sociotechnical systems thinking, safety engineering, biomedical ethics, dignity theory, human-centered AI, and governance-by-design.

The chapter proceeds in five stages. First, it defines the principal terms used throughout the thesis. Second, it explains the theoretical foundations that justify the framework's design choices. Third, it offers a formal academic definition of NAI 2.0™ as a governance architecture. Fourth, it examines the framework's four constitutional components---3ZEROS Sanctuary, Sacred Pause™, Sovereign Brake, and Tiger .1x Key™---and explains how they function together to constrain unsafe autonomy and preserve meaningful human authority. Finally, it introduces dignity as a system-level requirement and positions the

Elder Dignity Score as an exploratory evaluative instrument.

The chapter does not claim that conceptual rigor alone is sufficient for clinical adoption or regulatory acceptance. Rather, it establishes the intellectual structure through which later chapters will analyse the framework's technical design, governance relevance, and feasibility. Its role is therefore foundational: it translates an initial design vision into a form that can be critically examined as a scholarly contribution.

## 3.2 Definitions of Key Concepts

### 3.2.1 Non-agentic AI

For the purposes of this thesis, non-agentic AI refers to an AI system that remains bounded within advisory, interpretive, or constrained decision-support functions and does not independently execute or finalize high-impact care actions. A non-agentic system may generate recommendations, classify patterns, surface risks, or support workflow prioritization, but its outputs do not by themselves constitute clinical action. Instead, they remain subject to explicit human review, contextual interpretation, and, where necessary, refusal or modification.

This definition is important because the term non-agentic is not being used here to imply technical simplicity, inferiority, or low capability. Rather, it denotes a governance position: the system is intentionally designed so that intelligence does not equate to delegated authority. In eldercare, this distinction is ethically and operationally significant. Older adults may experience fluctuating capacity, dependency, communication barriers, or heightened vulnerability to harm arising from misclassification, neglect, coercion, or automation bias. Under such conditions, limiting AI authority is not a deficiency but a protective design choice.

### 3.2.2 Agentic AI

In contrast, agentic AI refers to systems capable of initiating, sequencing, or executing meaningful actions with reduced or delayed human intervention. Agentic systems are characterized not merely by analytical power but by a degree of operational authority: they can act, or materially shape action, in ways that exceed passive recommendation. In some sectors such systems may be desirable where speed, automation, and autonomous adaptation are primary goals. However, in eldercare settings involving safety-critical and dignity-sensitive decisions, such architectures may introduce unacceptable risks if they reduce the practical ability of humans to interrupt, contextualize, or reject machine outputs.

The distinction between agentic and non-agentic AI is therefore central to the thesis. The question is not whether AI can act more independently, but whether it should do so in a context where care is relational, morally loaded, and often irreversible in consequence.

### 3.2.3 Sovereign governance

Sovereign governance, as used in this thesis, refers to a model of AI oversight in which ultimate authority over high-impact care decisions remains with accountable human actors and is structurally protected from erosion by system design. Sovereign governance is not identical to general human oversight. Rather, it emphasizes the preservation of final decisional authority, interruption rights, and normative control even when AI systems are integrated deeply into workflow.

The use of the term sovereign is deliberate. It signals that human authority must not be merely symbolic or procedural. In many practical deployments, a human may appear to remain "in the loop" while in reality operating under time pressure, interface bias, institutional incentives, or design constraints that make deviation from AI output unlikely.

Sovereign governance requires more than nominal participation. It requires an architecture in which the human role is materially empowered, not performatively retained.

### 3.2.4 Human sovereignty

Human sovereignty refers to the preservation of meaningful human authority over high-impact decisions that affect the welfare, dignity, bodily integrity, privacy, or care trajectory of older adults. The term encompasses more than the ability to click "approve" or "decline." It includes the practical capacity to pause action, question the system, request additional context, invoke an override, and interpret recommendations against the lived circumstances of the person receiving care.

Human sovereignty is especially important in eldercare because decisions may involve subtle judgments about vulnerability, consent, family context, cultural norms, emotional state, or fluctuating clinical presentation. These are domains in which machine inference may be informative but cannot be presumed sufficient. A sovereign framework therefore insists that the system remain subordinate to human judgment in precisely those situations where consequences are greatest and values are most contested.

### 3.2.5 Constitutional architecture

A constitutional architecture is a system design in which normative limits are embedded structurally rather than expressed only in policy documents or user guidance. The term is used analogically. Just as a constitutional order defines the powers, limits, and procedures that govern legitimate action, a constitutional AI architecture defines what the system may do, what it may not do, what requires escalation, and what must remain under explicit human control.

Within this thesis, constitutional architecture does not imply legal constitutionalism in a formal state sense. Instead, it denotes rule-bounded design with enforceable constraints. The value of such an approach is that it shifts governance from aspiration to operation. Rather than trusting downstream users to remember ethical principles under pressure, the architecture itself imposes conditions on system behavior.

### 3.2.6 Hardware-enforced constraint

A hardware-enforced constraint is a boundary on system operation that is implemented through physical, device-level, or tightly coupled infrastructural controls rather than through software rules alone. The significance of this distinction lies in robustness. Software safeguards can be altered, ignored, bypassed, or weakened under conditions of urgency, poor implementation, or organizational drift. Hardware-enforced constraints are intended to make certain actions more difficult, slower, or impossible unless specified human conditions are met.

In this thesis, hardware enforcement is not treated as a magical solution to governance failure. It is instead understood as an additional layer of protection that supports interruptibility, bounded activation, and non-bypassable human authorization in high-impact situations. Its importance lies in shifting oversight from recommendation to enforcement.

### 3.2.7 Dignity preservation

Dignity preservation refers to the protection of the older person's status as a person who should be treated with respect, explanation, non-coercion, privacy, and meaningful recognition within care processes. In this thesis, dignity is not reduced to a single abstract moral claim. It is treated as a multidimensional concern that may be affected by how a system structures pace, permission, refusal, explanation, privacy, and human presence.

Dignity preservation therefore has both ethical and operational meaning. A system may be highly accurate yet still dignity-undermining if it accelerates decisions without explanation, constrains refusal, intensifies surveillance beyond necessity, or replaces relational care with impersonal optimization. The thesis argues that dignity should be treated as a design concern rather than an afterthought.

### 3.2.8 High-impact care decision

A high-impact care decision is any decision, escalation, intervention, or workflow progression that could materially affect an older person's health status, safety, autonomy, privacy, care access, or lived experience of dignity. Such decisions include, but are not limited to, urgent escalation, changes in care pathway, behavioral interpretation with restrictive consequences, privacy-sensitive monitoring shifts, or recommendations that may alter treatment attention or resource allocation.

The category matters because not all outputs require the same level of control. A bounded framework should distinguish between low-impact support functions and outputs that warrant stronger human review. This thesis therefore ties governance intensity to impact level rather than applying a uniform control model to all system outputs.

### 3.2.9 Meaningful human oversight

Meaningful human oversight refers to a form of human involvement that is informed, empowered, and operationally effective. It differs from nominal oversight in three ways. First, the human must have sufficient information to understand the significance of the system output. Second, the human must have real authority to interrupt or reject the workflow. Third, the timing of oversight must be early enough to matter. Oversight that occurs only after an action has already shaped care outcomes may satisfy formal procedure while failing substantively.

Meaningful oversight is thus inseparable from design. A human cannot exercise real control if the system is opaque, the interface is coercive, the workflow is rushed, or the override is practically inaccessible. The framework developed here assumes that oversight must be architecturally supported if it is to remain meaningful in real settings.

### 3.2.10 Preventable harm

Preventable harm refers to avoidable injury, deterioration, distress, rights violation, or dignity erosion that could reasonably have been reduced through safer design, better oversight, or more appropriate escalation structure. The term does not imply that all adverse outcomes are foreseeable or that care systems can eliminate risk entirely. Rather, it recognizes that some harms arise not only from human error or model weakness, but from poor governance architecture.

In this thesis, the relevance of preventable harm lies in design responsibility. If a system allows high-impact outputs to progress without review, or if it normalizes overreliance on automated prompts, then some resulting harms may be governance failures rather than isolated mistakes. A non-agentic constitutional design is therefore proposed as a mechanism for reducing preventable harm by limiting the conditions under which unsafe progression can occur.

## 3.3 Theoretical Foundations

### 3.3.1 Sociotechnical systems theory

The first major theoretical foundation of this thesis is

sociotechnical systems theory, which holds that technologies cannot be understood in isolation from the human, institutional, and organizational environments in which they operate. In healthcare, and especially in eldercare, AI does not function as an independent object. It becomes part of a wider network involving caregivers, clinicians, administrators, family members, routines, documentation practices, risk perceptions, staffing pressures, and care values.

This perspective is critical because it prevents the framework from being reduced to model performance alone. A technically impressive system may still be unsafe if it is embedded in an environment where users are overburdened, escalation pathways are unclear, or interface design encourages compliance without reflection. Conversely, a bounded system may improve safety not because its model is more powerful, but because its governance architecture fits the realities of care practice more responsibly. Sociotechnical theory therefore supports the thesis's claim that eldercare AI should be designed with attention to workflow, authority, interruptibility, and human context, not merely algorithmic output.

### 3.3.2 Safety engineering and fail-safe design

A second foundation is drawn from safety engineering, particularly principles relating to fail-safe design, hazard containment, and interruptibility. Safety engineering assumes that complex systems will face uncertainty, incomplete information, and the possibility of error. In such settings, the goal is not to assume flawless operation, but to ensure that failures do not escalate uncontrollably and that the system remains governable under stress.

Applied to eldercare AI, this perspective supports several core features of NAI 2.0™. First, high-impact outputs should trigger additional review rather than seamless progression. Second, interruption should be easy, available, and authoritative. Third, no single inference should silently exceed its intended operational role. Fourth, logging and auditability matter because they allow later examination of near misses, deviations, and system-induced vulnerabilities.

Importantly, safety engineering also justifies the distinction between software safeguards and hardware-enforced controls. If the consequences of a bypass are serious, then relying solely on user compliance or interface reminders may be insufficient. The framework's emphasis on hard constraints is therefore consistent with a safety culture that assumes pressure, drift, and fallibility as normal rather than exceptional.

### 3.3.3 Biomedical ethics and dignity theory

The third theoretical foundation comes from biomedical ethics, especially the principles of respect for persons, non-maleficence, beneficence, and justice, alongside more specific work on dignity in healthcare. In eldercare, these principles acquire particular urgency because the persons affected may experience dependency, frailty, reduced bargaining power, or forms of social invisibility that make them more susceptible to paternalism or procedural harm.

This thesis places special emphasis on dignity because it captures dimensions of care that are often not adequately represented by safety and efficiency metrics alone. A system may avoid obvious physical harm while still compromising a person's dignity by accelerating decisions without explanation, eroding the right to refuse, or subjecting them to forms of surveillance that are technically permissible but relationally degrading. Dignity theory thus deepens the governance argument by requiring that care technologies preserve not only life and function, but personhood and relational respect.

From this perspective, sovereignty and dignity are linked. To preserve dignity is partly to preserve the conditions under which the person is not reduced to a data object or workflow variable. Human review, explanation, pause, and context-sensitive judgment are therefore not merely administrative features; they are ethically meaningful practices.

### 3.3.4 Human-centered AI and governance-by-design

The fourth foundation is human-centered AI, combined with the more specific concept of governance-by-design. Human-centered AI emphasizes systems that support, rather than supplant, human values, capacities, and interpretive roles. Governance-by-design extends this logic by arguing that ethical and regulatory principles should be built into architecture rather than left entirely to training or downstream enforcement.

This combined foundation is especially relevant to the thesis because it explains why the framework is organized around boundaries, authorizations, and pauses rather than around predictive expansion. NAI 2.0™ is not primarily a capability-maximizing system. It is a capability-bounding system. Its purpose is to support care without displacing accountability. Governance-by-design makes this possible by embedding normative expectations---such as interruptibility, review, role-sensitive access, and logging---directly into the structure of operation.

Together, these theoretical foundations justify the core design philosophy of the thesis: eldercare AI should be treated as a governable sociotechnical participant in care, bounded by safety logic, ethically responsive to dignity, and architecturally configured to preserve human authority.

## 3.4 Formal Definition of NAI 2.0™

NAI 2.0™ is defined in this thesis as a hardware-enforced non-agentic AI governance framework for high-impact eldercare environments, designed to preserve meaningful human sovereignty over care decisions while supporting safety, accountability, privacy-conscious operation, and dignity-preserving practice.

This definition has five implications.

First, NAI 2.0™ is a governance framework, not merely a model or application. Its primary contribution lies in how authority is structured, constrained, and reviewed. Second, it is non-agentic: the system may generate bounded outputs, but it does not autonomously execute high-impact actions. Third, it is hardware-enforced in the sense that critical constraints are not left entirely to software discretion or user goodwill; rather, the architecture incorporates stronger interruption and authorization conditions intended to prevent unsafe progression. Fourth, it is eldercare-specific, meaning that its design is responsive to the relational, ethical, and vulnerability features of care for older adults. Fifth, it is constitutionally organized, in that system behavior is shaped by predefined normative limits on authority, escalation, and permissible action.

The purpose of NAI 2.0™ is therefore not to replace professional judgment, family involvement, or care relationships. Nor is it to eliminate all uncertainty from care. Instead, it is to create an operational environment in which AI can assist without assuming a role that exceeds what is ethically and clinically appropriate. The framework assumes that high-impact care settings require not only intelligent tools but also disciplined structures of restraint.

In this sense, NAI 2.0™ can be understood as a bounded constitutional layer around AI-supported eldercare. It seeks to ensure that machine-generated outputs remain reviewable, interruptible, and subordinate to accountable human actors. The framework thereby treats governance not as an external audit function alone, but as an internal feature of system design.

## 3.5 Core Constitutional Components

### 3.5.1 3ZEROS Sanctuary

3ZEROS Sanctuary is defined as a bounded operational environment within which AI-mediated eldercare interaction is governed by explicit constraints on unsafe autonomy, unauthorized progression, and unreviewed high-impact output. The concept of sanctuary is used here to denote a protected domain of care in which the system is not permitted to exceed predefined limits of authority.

Academically understood, 3ZEROS Sanctuary serves three functions. First, it establishes boundary conditions: what kinds of actions the AI may support, what kinds it may not initiate, and what categories of output automatically require additional human scrutiny. Second, it creates a

protected mode of operation in which high-impact transitions cannot occur silently or informally. Third, it embodies a care-centered philosophy by treating safety and dignity as conditions of entry into system action rather than optional considerations after the fact.

The value of this component lies in its environmental logic. Rather than relying solely on individual users to remember when caution is needed, the sanctuary concept frames the whole operating context as one in which boundedness is assumed. This reduces the risk that efficiency pressures or routine normalization will gradually expand the system's role beyond what is appropriate. In eldercare, where subtle shifts in dependency or distress can carry serious implications, such containment is a practical form of ethical discipline.

### 3.5.2 Sacred Pause™

Sacred Pause™ refers to a mandatory interruption layer that suspends progression from AI output to consequential action whenever predefined thresholds of impact, uncertainty, or dignity relevance are reached. The word sacred does not carry theological meaning in this thesis. It is used to signal non-triviality: the pause is not a superficial delay but a protected moment of human review that the workflow is not permitted to bypass casually.

The concept addresses a central problem in AI-assisted care: the tendency for recommendations to gain practical authority through speed, interface design, or workflow

momentum. Even when a human is technically present, decisions may unfold too quickly for reflection. Sacred Pause™ is intended to counter this by making review structurally unavoidable at the point where it matters most.

The pause performs at least four governance functions. It creates time for contextual interpretation; it reduces automation bias by interrupting seamless progression; it restores room for explanation and discussion; and it provides a procedural space in which the older person's dignity-related interests may still be considered before action crystallizes. In high-impact eldercare, delay is not always a defect. In some circumstances, it is a safeguard against ethically and clinically premature action.

### 3.5.3 Sovereign Brake

Sovereign Brake is the framework's immediate override and halt mechanism. It exists to ensure that any authorized human actor can stop workflow progression when safety, consent, contextual uncertainty, or dignity concerns make continuation inappropriate. The brake is sovereign because it reflects retained human supremacy over the system's trajectory.

This mechanism is conceptually distinct from ordinary review. Review assesses whether an output should proceed; the brake ensures that the answer can be "no" in a definitive and operationally effective way. It is therefore a crucial component of meaningful human oversight. If a system can be reviewed but not truly stopped, then oversight is weakened. The brake restores decisional asymmetry in favor of the human.

In eldercare settings, this matters because circumstances are often dynamic and ambiguous. A recommendation that appears sensible at the level of data may become inappropriate when interpreted in the context of patient distress, family preference, cultural sensitivity, or bedside observation. Sovereign Brake provides the structural means to privilege such contextual knowledge over system momentum.

### 3.5.4 Tiger .1x Key™

Tiger .1x Key™ is defined as a bounded authorization mechanism through which particular functions, thresholds, or workflow progressions require validated human permission before activation. Its purpose is to ensure that certain transitions cannot occur merely because the system produced an output. Instead, progression depends upon explicit, role-sensitive authorization.

The key contributes to governance in three principal ways. First, it supports permission control, ensuring that authority is not diffused ambiguously across users or system states. Second, it supports

traceability, because authorization events can be logged and reviewed. Third, it reinforces non-agentic design by making activation contingent upon human validation rather than internal AI confidence alone.

The importance of this component lies in recognizing that not all human involvement is equivalent. A bounded authorization mechanism requires the right kind of human involvement at the right point in the workflow. In this sense, Tiger .1x Key™ operationalizes sovereignty by linking power to explicit permission rather than passive observation.

## 3.6 Dignity as a System-Level Requirement

A central claim of this thesis is that dignity should be treated as a system-level requirement, not merely as a background ethical aspiration. This does not mean that dignity can be captured exhaustively by formal rules. Dignity remains context-sensitive, relational, and partly resistant to quantification. However, it does mean that system design can either support or undermine conditions commonly associated with dignified care.

Within eldercare, dignity may be affected by whether a person is given understandable explanation, whether there is time to question or refuse a recommendation, whether privacy is protected proportionately, whether care remains personalized, and whether machine mediation displaces human presence in moments of significance. A framework that ignores these factors may be efficient yet degrading. By contrast, a system that preserves pause, explanation, refusal, and contextual review is more likely to support dignified care even if it does not solve every ethical challenge.

Treating dignity as a system-level requirement therefore expands the evaluative scope of AI governance. It asks not only whether the system is accurate or compliant, but whether it helps preserve the personhood of those subjected to its outputs. In this thesis, dignity is not opposed to safety; rather, it is part of what safe and ethical eldercare should mean.

## 3.7 The Elder Dignity Score

The Elder Dignity Score is introduced in this thesis as an

exploratory evaluative instrument intended to operationalize selected dignity-related dimensions of AI-supported eldercare. It is not presented as a universally validated clinical scale at this stage. Instead, it functions as a structured means of examining whether system design and workflow appear to preserve important elements of dignity, such as perceived control, privacy, explanation quality, non-coercion, emotional comfort, and adequacy of human presence.

Its inclusion reflects the broader theoretical argument of the chapter: if dignity matters in eldercare AI, then it should be visible in evaluation as well as in rhetoric. At the same time, the instrument is used cautiously. Dignity cannot be fully reduced to a score, and any such

measure must be interpreted within wider qualitative and contextual understanding. The value of the Elder Dignity Score is therefore heuristic and developmental. It helps make dignity discussable, reviewable, and improvable within system design without claiming to exhaust the moral richness of dignified care.

## Concluding Note to Chapter 3

This chapter has defined the conceptual foundations of the thesis and formalized NAI 2.0™ as a hardware-enforced non-agentic governance framework for eldercare. It has argued that the key problem is not simply whether AI can perform useful analytical tasks, but how its authority is bounded in contexts where older adults may be vulnerable to both clinical and dignity-related harms. By defining non-agentic design, sovereign governance, constitutional architecture, hardware-enforced constraints, and dignity preservation in precise terms, the chapter provides the conceptual basis for the rest of the thesis.

The theoretical discussion has shown that the framework is supported by sociotechnical, safety, ethical, and governance-by-design traditions rather than by a single disciplinary logic. It has also clarified the role of the four constitutional components and positioned dignity as an operational requirement rather than a secondary aspiration. With these foundations established, the next chapter can move from conceptual definition to methodological design, explaining how the framework is evaluated and what forms of evidence are used to assess its coherence, feasibility, and governance relevance.

# Chapter 4. Methodology

## 4.1 Introduction

This chapter explains the methodological approach used to develop and evaluate NAI 2.0™ as a hardware-enforced non-agentic AI governance framework for eldercare. Its purpose is to show how the thesis moves from a conceptual argument to a structured research process capable of supporting defensible claims. Because the thesis is concerned not only with technical design, but also with governance, safety, dignity, and regulatory alignment, no single disciplinary method is sufficient on its own. The study therefore adopts an interdisciplinary methodology that combines design science research, conceptual analysis,

regulatory mapping, and feasibility-oriented evaluation.

The central methodological challenge of the thesis lies in the nature of the research problem itself. The work does not ask only whether an AI model can perform a specific predictive task. Rather, it asks how an AI system should be designed, constrained, and governed in a high-impact eldercare context so that meaningful human control, dignity preservation, safety, and compliance-readiness are maintained. This means that the research must address technical architecture, normative reasoning, workflow design, and institutional governance together. A purely experimental approach would be too narrow, while a purely theoretical approach would be insufficiently grounded in deployment realities. The methodology must therefore support both artefact creation and critical evaluation.

For this reason, the thesis adopts a design-oriented strategy in which the primary research output is a governance artefact: the NAI 2.0™ framework. The methodological logic is iterative. The framework is derived from problem identification, informed by literature across healthcare AI, eldercare ethics, safety engineering, and regulatory governance, and then evaluated through structured analytical and feasibility-oriented lenses. The study does not claim to present a completed large-scale clinical effectiveness trial. Instead, it offers a rigorous design-and-evaluation approach appropriate to an emerging governance architecture at this stage of development.

This chapter proceeds in eight sections. Section 4.2 explains the overall research design and justifies the choice of design science research. Section 4.3 outlines the phases of the study from problem identification through framework refinement. Section 4.4 describes the data sources and materials used. Section 4.5 defines the outcome domains used to evaluate the framework. Section 4.6 explains the role and structure of the Elder Dignity Score as an exploratory assessment tool. Section 4.7 addresses research ethics, particularly in relation to vulnerable populations and privacy-sensitive care contexts. Section 4.8 acknowledges the methodological limitations of the thesis. Together, these sections establish the basis on which later claims about architecture, governance, and feasibility should be interpreted.

## 4.2 Research Design

The study adopts a design science research orientation because its main objective is to create and evaluate an artefact intended to address a real-world problem. In this thesis, that artefact is not merely a software application or a single predictive model. It is a governance architecture for eldercare AI deployment: NAI 2.0™. Design science research is appropriate in such circumstances because it is concerned with the purposeful development of artefacts that solve identified problems while also contributing to knowledge. The approach is particularly well suited to research situated between technical systems design and institutional governance, where the aim is neither abstract theorization alone nor simple empirical observation of existing practice.

The use of design science research rests on three considerations. First, the thesis addresses a practical problem: the absence of an enforceable and eldercare-specific framework for bounded AI deployment in high-impact care settings. Second, the study seeks to generate a constructive response rather than merely critique current systems. Third, the resulting framework must be evaluated not only for conceptual coherence but also for practical plausibility, governance alignment, and ethical defensibility. Design science allows these aims to be integrated within a single methodological structure.

However, design science research alone does not fully capture the needs of this thesis. The problem of eldercare AI governance is also legal, ethical, and sociotechnical. For that reason, the study is supplemented by three additional methodological components.

The first is conceptual analysis. This is necessary because the thesis relies on terms such as non-agentic AI, human sovereignty, constitutional architecture, and dignity preservation that are used inconsistently in current literature. The study therefore develops precise thesis-specific definitions and clarifies the normative assumptions embedded within the framework. Conceptual analysis is not treated here as detached philosophical exercise; rather, it serves to ensure that the artefact is analytically coherent and that later evaluation is conducted against explicit criteria.

The second component is regulatory and governance mapping. Since a central claim of the thesis is that NAI 2.0™ is compliance-supportive and governance-ready, the framework must be evaluated against relevant expectations from healthcare AI governance, Software as a Medical Device reasoning, data protection principles, and high-risk AI oversight models. This does not amount to legal certification. Instead, it is a structured analytical process through which the artefact is compared with externally defined governance requirements and assessed for alignment, tension, or unresolved gaps.

The third component is feasibility-oriented evaluation. Because the framework is meant for real care settings, the study must examine not only whether the design is theoretically persuasive but also whether it is operationally plausible. Feasibility in this thesis refers to the practicality of workflow integration, the intelligibility of control mechanisms, the viability of human review points, and the potential acceptability of the framework to relevant stakeholders. Depending on the available evidence base, this may be assessed through

protocol design, simulation logic, pilot workflow analysis, expert review, scenario testing, or preliminary implementation observations.

This mixed methodological strategy is justified by the nature of the research questions. The primary question concerns how hardware-enforced non-agentic AI can preserve human sovereignty, dignity, and safety in eldercare while supporting regulatory alignment. Such a question cannot be answered solely by quantitative metrics or purely doctrinal analysis. It requires a method that can design an artefact, justify its normative basis, map it against governance expectations, and examine its operational plausibility.

In methodological terms, the study is therefore constructive,

analytical, and evaluative. It is constructive because it builds a new framework. It is analytical because it clarifies and integrates concepts from multiple fields. It is evaluative because it tests the framework against defined criteria rather than presenting it uncritically. This multi-layered design strengthens the credibility of the thesis by ensuring that NAI 2.0™ is not offered as a rhetorical proposition but as a reasoned and assessable governance model.

## 4.3 Study Phases

The study was organized into five main phases. These phases are presented sequentially for clarity, although some iteration occurred between them as the framework was refined in response to emerging analysis.

### 4.3.1 Phase 1: Problem Identification and Framing

The first phase involved identifying and framing the central research problem. This phase began from the observation that AI deployment in healthcare, including eldercare, is often discussed in terms of innovation, efficiency, and prediction, while the question of bounded authority remains insufficiently addressed. In eldercare, where vulnerable persons may be affected by high-impact recommendations, there is a particular risk that machine output will gain practical authority without adequate structural oversight.

Problem identification was informed by early reading across healthcare AI, eldercare technology, dignity ethics, patient safety, and AI governance. The aim at this stage was not yet to design a solution, but to clarify what exactly was missing in current approaches. Three interrelated deficiencies were identified. First, many existing AI governance discussions focus on principles without showing how those principles are embedded in system architecture. Second, eldercare-specific concerns such as dependency, dignity, and relational care are often treated as secondary to performance or operational efficiency. Third, human oversight is frequently invoked in formal terms without sufficient attention to whether it is meaningful, timely, or structurally enforceable.

This phase led to the formulation of the thesis problem as the lack of a governance architecture that can meaningfully constrain AI authority in high-impact eldercare settings while preserving human sovereignty and dignity. It also informed the final research questions and justified the move toward a design-oriented methodology.

### 4.3.2 Phase 2: Literature-Informed Conceptual Development

The second phase involved conceptual development grounded in a structured review of relevant literature. This phase served two purposes. First, it established the scholarly basis of the thesis by situating the work in relation to existing debates. Second, it provided the conceptual resources needed to define the framework.

Literature from several domains was examined: AI in healthcare, eldercare deployment, human oversight, meaningful human control, safety engineering, dignity theory, biomedical ethics, Software as a Medical Device governance, privacy regulation, and constitutional or rule-bounded AI systems. Rather than treating these literatures as isolated, the study integrated them around the common problem of bounded and accountable AI deployment.

During this phase, the thesis developed working definitions of the key constructs used throughout the study, including non-agentic AI,

human sovereignty, constitutional architecture,

hardware-enforced constraint, and dignity preservation. These definitions were refined iteratively as the framework took shape. The literature also informed the identification of the four core constitutional components of NAI 2.0™: 3ZEROS Sanctuary, Sacred Pause™, Sovereign Brake, and Tiger .1x Key™.

This phase was not a neutral summary exercise. It was interpretive and synthetic. The purpose was to identify where current literature leaves unresolved tensions and how a new governance architecture might respond to them.

### 4.3.3 Phase 3: Artefact Design and Architecture Formation

The third phase focused on designing the governance artefact itself. Here, the conceptual insights from the literature were translated into an operational architecture. The framework was developed as a layered model in which data intake, inference, constitutional constraints, hardware enforcement, human review, logging, and fallback procedures were specified as interdependent governance components.

The design process involved several core questions:

— What should the AI system be permitted to do?

— What must remain under human authority?

— Which outputs should trigger pause, review, or override?

— How can oversight be made operational rather than symbolic?

— How can privacy and dignity be reflected in workflow design?

— What architectural features are needed to support traceability and regulatory alignment?

Through this phase, NAI 2.0™ was formalized as a bounded decision-support architecture rather than an autonomous action system. The design also incorporated distinctions between low-impact, moderate-impact, and high-impact outputs, allowing different governance intensities to be attached to different classes of system behavior.

This phase produced the core material later presented in Chapter 5. Importantly, the artefact was not designed in purely technical terms. Governance, ethics, and workflow considerations were treated as constitutive features of the system rather than external constraints applied after design.

### 4.3.4 Phase 4: Analytical Evaluation and Mapping

The fourth phase consisted of structured analytical evaluation. In this phase, the framework was assessed against defined criteria drawn from regulation, ethics, safety logic, and implementation feasibility. The purpose was to test whether the artefact, as designed, could plausibly support the claims made for it.

This phase included:

— regulatory mapping against healthcare AI governance expectations

— privacy and data governance analysis in relation to key data protection principles

— ethical analysis focused on autonomy, dignity, non-maleficence, and accountability

— workflow evaluation examining whether pause and authorization logic could plausibly function in practice

— risk review considering failure modes such as over-reliance, alert fatigue, delay, and implementation drift

This phase did not produce a binary result of "valid" or "invalid." Rather, it generated a structured understanding of where the framework is strong, where it is only partially supported, and where further evidence is needed. This is methodologically important because the thesis does not overclaim. The evaluation phase distinguishes between conceptual coherence, architectural plausibility, governance alignment, and empirical proof.

### 4.3.5 Phase 5: Feasibility Reflection and Refinement

The fifth and final phase involved reflecting on the outputs of the evaluation and refining the framework accordingly. This phase recognized that governance artefacts are rarely final on first design. Questions of burden, timing, implementation complexity, and dignity measurement required further adjustment and qualification.

Where feasibility evidence was available in pilot, simulation, protocol, or expert-assessment form, it was used to identify practical strengths and tensions. Where such evidence remained prospective, the thesis articulated these elements as validation pathways rather than completed proof. This phase also informed the positioning of the Elder Dignity Score as an exploratory evaluative tool rather than a fully validated instrument.

The final result of this phase was a revised and more disciplined articulation of the thesis claims. The framework was presented not as a finalized commercial product, but as a robust governance architecture whose conceptual, architectural, and regulatory foundations are stronger than its current large-scale empirical evidence base. This distinction is carried throughout the remainder of the thesis.

## 4.4 Data Sources

The thesis draws on multiple sources of material because the research problem spans design, governance, ethics, and deployment. No single source type would be adequate to assess the framework fully. The methodological strategy therefore uses a multi-source evidence base appropriate to interdisciplinary design research.

The first major source category is scholarly literature. This includes peer-reviewed work on healthcare AI, eldercare technology, AI governance, human oversight, safety engineering, biomedical ethics, dignity theory, and privacy regulation. These materials were used to identify the research gap, define key concepts, inform the artefact design, and support the analytical arguments made throughout the thesis. Literature served not only as background but as a core source of design requirements and evaluative criteria.

The second category is regulatory and policy documentation. This includes materials relevant to healthcare software governance, high-risk AI regulation, and privacy law. Such documents were used to derive governance expectations against which NAI 2.0™ could be mapped. The study does not treat regulatory texts as static or universally transferable. Instead, they are used as structured reference frameworks for analytical comparison.

The third category is system design material generated by the thesis itself. This includes architectural diagrams, workflow definitions, component descriptions, escalation logic, risk matrices, and conceptual design documents associated with NAI 2.0™. These artefacts are not secondary illustrations; they are primary research outputs within a design science methodology and therefore legitimate objects of analysis.

The fourth category, where available, consists of feasibility-oriented materials. Depending on the evidentiary basis of the project, this may include pilot protocol designs, simulation records, structured scenarios, stakeholder observations, early implementation notes, expert feedback, or workflow assessments. These materials are used cautiously. They are not presented as definitive clinical outcome evidence but as practical inputs into feasibility analysis.

The fifth category includes derived evaluative instruments, particularly the Elder Dignity Score in exploratory form. Although not yet a validated psychometric measure, it constitutes a structured source of evaluative information regarding dignity-related aspects of system design and workflow.

This multi-source approach is methodologically justified because the thesis evaluates a governance architecture rather than a single predictive endpoint. A framework concerned with authority, interruption, dignity, and compliance-readiness must be assessed using evidence that speaks to technical design, normative reasoning, and deployment practicality together.

## 4.5 Outcome Measures

Because NAI 2.0™ is a governance framework rather than a standalone clinical prediction model, its evaluation requires multiple outcome domains. The thesis therefore does not rely on a single primary endpoint. Instead, it distinguishes between safety-related,

oversight-related, feasibility-related, dignity-related, and

governance-related outcomes. This approach reflects the broader argument that eldercare AI should be evaluated not only by predictive performance, but by how it structures authority and care.

### 4.5.1 Safety-Related Outcomes

Safety-related outcomes concern whether the architecture plausibly reduces the risk of preventable harm arising from unreviewed or inappropriate AI-supported action. In a full empirical deployment, such outcomes might include the frequency of unsafe progression events, missed escalation, inappropriate automated reliance, or breakdown in interruption logic. At the current stage of the thesis, safety is assessed more modestly through architectural and workflow criteria such as:

- whether high-impact outputs are prevented from self-executing
- whether clear interruption mechanisms exist
- whether fallback pathways are defined
- whether known failure modes are anticipated in design
- whether logging supports investigation of adverse events or near misses

These measures are appropriate to a design-stage study because they evaluate whether safety has been structurally considered rather than assumed.

### 4.5.2 Oversight-Related Outcomes

Oversight-related outcomes assess whether human control is meaningful in operational terms. These include:

- the number and location of human review points
- whether review occurs before consequential progression
- whether override functions are practically available
- whether authorization is explicit and attributable
- whether the interface supports informed rather than symbolic review

These outcomes are central to the thesis because meaningful human control is one of its core claims. A system cannot be described as non-agentic in any serious sense if human oversight is nominal, late, or structurally weak.

### 4.5.3 Feasibility-Related Outcomes

Feasibility outcomes concern the practical usability and workflow compatibility of the framework. These include:

— whether pause and review logic can plausibly fit within eldercare workflows

— whether the escalation structure is understandable

— whether hardware-enforced controls are operationally realistic

— whether the system imposes manageable rather than excessive burden

— whether the framework is acceptable in principle to relevant users or stakeholders

At the thesis stage, these outcomes may be assessed through structured scenario analysis, pilot protocol logic, expert review, or preliminary implementation observations. The emphasis is on practical plausibility rather than large-scale deployment success.

### 4.5.4 Dignity-Related Outcomes

Dignity-related outcomes address the extent to which the framework appears to preserve conditions associated with dignified care. Because dignity is multidimensional and partly qualitative, these outcomes are interpreted cautiously. They may include:

— perceived control over care-related decisions

— opportunity for explanation and pause

— non-coercive workflow structure

— proportional privacy protection

— adequacy of human presence and review

— avoidance of depersonalizing automation

These dimensions are examined partly through qualitative reasoning and partly through the exploratory Elder Dignity Score described below.

### 4.5.5 Governance Alignment Outcomes

Governance outcomes concern the degree to which the framework aligns with relevant regulatory and institutional expectations. These include:

- clarity of intended use
- traceability of key actions
- role-based access and authorization
- support for lifecycle governance
- privacy-conscious data architecture
- compatibility with risk-based oversight models

These outcomes do not demonstrate formal certification. Rather, they provide structured evidence for the claim that the framework is governance-ready in design.

Taken together, these outcome domains allow the thesis to evaluate the framework in a manner appropriate to its nature. They also help maintain analytical honesty by distinguishing what can be assessed at the design-and-feasibility stage from what would require larger-scale empirical validation.

## 4.6 Elder Dignity Score Method

The Elder Dignity Score is introduced in this thesis as an

exploratory assessment tool intended to capture selected dignity-related dimensions of AI-supported eldercare. Its purpose is not to reduce dignity to a single number or to claim psychometric finality. Rather, it provides a structured way of examining whether the framework supports or undermines important features of dignified care.

The tool is organized around several dimensions that are especially relevant to eldercare in AI-mediated settings. These include:

- perceived control, referring to whether the person or their

representatives retain meaningful influence over consequential decisions

— explanation quality, referring to whether the reasons for escalation or review can be communicated intelligibly

— privacy respect, referring to whether data use and monitoring are proportionate

— non-coercion, referring to whether system-supported workflow leaves room for refusal, pause, or reconsideration

— relational respect, referring to whether human involvement remains visible and meaningful

— emotional comfort or reassurance, referring to whether the care process appears supportive rather than machine-dominant

In methodological terms, the score may be applied through structured observer assessment, stakeholder review, simulated case assessment, or mixed evaluation depending on the evidence available. At this stage of the thesis, it is best understood as a developmental instrument rather than a finalized scale. Where ratings are used, they are interpreted comparatively and cautiously rather than as clinically decisive outputs.

The reason for including this tool is methodological as well as ethical. If dignity is treated as a central design goal, then the study requires some structured means of evaluating it. Without such a tool, dignity risks remaining rhetorical. The Elder Dignity Score helps make dignity visible within design assessment while still acknowledging its limitations. The thesis therefore uses it as a supplementary evaluative mechanism, not as a substitute for broader ethical judgment or qualitative interpretation.

## 4.7 Research Ethics

The research raises significant ethical considerations because it concerns AI deployment in eldercare, a domain involving potentially vulnerable individuals, sensitive personal information, and high-impact decisions affecting welfare and dignity. Even where the thesis is design-oriented and does not itself constitute a large-scale clinical intervention study, ethical reflection remains essential.

The first major ethical consideration is vulnerability. Older adults may experience frailty, cognitive fluctuation, communication barriers, dependency on caregivers, or reduced ability to contest system-supported decisions. For this reason, the framework developed in the thesis is intentionally non-agentic and preserves human review and interruption rights. This is not only a design choice but an ethical response to the possibility that more autonomous systems could intensify existing asymmetries of power.

The second concern is informed consent and participation, where applicable. If any empirical component of the project involves human participants, whether through pilot observation, stakeholder interviews, workflow assessment, or early testing, informed consent procedures must be proportionate to the participant group and sensitive to issues of capacity, representation, and voluntariness. No participant should be exposed to a system that replaces professional judgment or creates unmanaged risk.

The third concern is privacy and data minimization. Because eldercare data may include health information, behavioral observations, and sensitive contextual details, the study adopts a minimum-necessary principle. Any data used for design, review, or feasibility assessment should be limited to what is needed for the research purpose, appropriately secured, and handled in a way that protects confidentiality. Where possible, de-identification or anonymization principles should be applied.

The fourth concern is non-maleficence. The framework should not be evaluated in a way that exposes individuals to unsafe automation, coercive workflow, or unjustified surveillance. This is particularly relevant if future pilot or trial phases are conducted. The thesis therefore positions many of its evaluation pathways as staged and bounded rather than assuming immediate large-scale live deployment.

The fifth concern is ethical review and governance approval. Where required by institutional policy, any human-participant or real-world evaluation phase should be subject to formal ethics review before implementation. At the thesis stage, conceptual development and architecture design may not in themselves require clinical ethics approval, but any transition into empirical deployment should be governed accordingly.

Overall, the research ethics position of this thesis is aligned with its substantive argument: eldercare AI should be designed and studied in ways that preserve dignity, minimize intrusion, avoid over-delegation, and maintain accountable human responsibility.

## 4.8 Methodological Limitations

Several methodological limitations should be acknowledged in order to interpret the thesis appropriately.

First, the study is design-oriented rather than a large-scale clinical effectiveness trial. Its primary contribution lies in the development and structured evaluation of a governance artefact, not in the production of definitive clinical outcome data across multiple institutions. This means that claims about safety improvement, workflow impact, or dignity preservation must remain proportionate to the evidence currently available.

Second, the thesis relies partly on conceptual and analytical evaluation rather than exclusively on live empirical testing. This is appropriate for an artefact at this stage of

maturity, but it also means that some conclusions are stronger at the level of architectural plausibility than at the level of demonstrated field effectiveness.

Third, the Elder Dignity Score remains exploratory. It is a useful developmental instrument, but it should not be treated as fully validated or psychometrically settled. Its findings, where used, must therefore be interpreted cautiously and in conjunction with broader ethical reasoning.

Fourth, the regulatory analysis presented in the thesis is

mapping-based rather than certifying. It identifies alignment with relevant governance principles but does not constitute formal legal advice, device classification, or jurisdiction-specific approval. Actual compliance would depend on institutional implementation, documentation, and regulatory engagement.

Fifth, the methodology is influenced by the fact that eldercare is a

highly variable sociotechnical environment. Care settings differ in staffing, resources, digital maturity, patient population, and legal context. As a result, transferability of the framework cannot be assumed without adaptation. The thesis offers a generalizable governance model, but not a claim of universal plug-and-play applicability.

Finally, there is the broader limitation that methodology in this field must contend with the dynamic nature of AI regulation and healthcare practice. Governance expectations are evolving, and design frameworks must be able to adapt accordingly. NAI 2.0™ should therefore be understood as a serious and structured contribution to an ongoing field of development rather than as a final settled solution.

These limitations do not undermine the thesis. Rather, they define its proper scope. The study makes its strongest contribution at the level of conceptual clarity, architectural design, governance integration, and feasibility-oriented reasoning. Larger empirical claims remain matters for future validation.

**Concluding Note to Chapter 4**

This chapter has outlined the methodological foundation of the thesis and explained how NAI 2.0™ was developed and assessed. The study adopts a design science research orientation because the central aim is to create and evaluate a governance artefact capable of addressing a real and under-specified problem in eldercare AI deployment. To support this aim, the methodology combines conceptual analysis, regulatory mapping, and feasibility-oriented evaluation within a structured multi-phase research process.

The chapter has also clarified the sources of evidence used in the study, the outcome domains by which the framework is assessed, and the role of the Elder Dignity Score as an exploratory evaluative instrument. In addition, it has shown that ethical reflection is not external to the research process but built into it, particularly in relation to vulnerable populations, privacy, and non-maleficence. Finally, it has acknowledged the study's

limitations, especially the distinction between design-stage evaluation and large-scale empirical proof.

This methodological clarity is essential for interpreting the rest of the thesis correctly. The framework should be read as a rigorously developed and critically assessed governance architecture whose claims are strongest at the conceptual, architectural, and compliance-supportive levels, while broader empirical claims remain appropriately bounded. With the methodology now established, the next chapter can turn to the technical substance of the artefact itself by presenting the system architecture and deployment model of NAI 2.0™ in full detail.

# Chapter 5. NAI 2.0™ System Architecture and Deployment Model

## 5.1 Introduction

This chapter sets out the technical and operational architecture of

NAI 2.0™ and explains how the framework translates the conceptual principles established in Chapter 3 into a bounded deployment model for eldercare. The central purpose of the chapter is not to present AI capability in isolation, but to show how system design can embody governance, oversight, safety, and dignity-preserving requirements in practice. In this thesis, the value of the proposed architecture lies less in autonomous technical performance than in its disciplined control of authority, escalation, and human decision rights.

The argument advanced here is that eldercare requires a deployment logic fundamentally different from capability-maximizing AI environments. Older adults often receive care within settings marked by frailty, fluctuating capacity, communication vulnerability, family involvement, privacy sensitivity, and high relational dependence. In such contexts, a technically strong system may still be inappropriate if it accelerates high-impact recommendations into action without sufficient review, explanation, or contextual interpretation. The architecture presented in this chapter is therefore deliberately designed around bounded operation. It assumes that the question is not only what the system can infer, but also what it must never be allowed to do without meaningful human authorization.

To make this argument concrete, the chapter proceeds in six stages. First, it defines the system boundaries that distinguish permissible support functions from prohibited or restricted forms of autonomous action. Second, it describes the main architectural layers through which data, inference, rule enforcement, authorization, and auditability are organized. Third, it illustrates the workflow of a representative care event to show how outputs move through the system in controlled rather than automatic ways. Fourth, it explains the rationale for hardware-enforced safety logic and clarifies why the framework does not rely solely on software guardrails or procedural policy. Fifth, it outlines the privacy and data handling assumptions that support a minimum-necessary and accountability-oriented deployment model. Sixth, it considers failure modes, risk controls, and the practical conditions required for implementation.

The chapter does not claim that architecture alone guarantees safe or ethical outcomes. No system design can eliminate the need for competent professionals, organizational training, institutional accountability, or context-sensitive judgment. Rather, the chapter argues that architectural design can materially improve governance conditions by narrowing the pathways through which unsafe autonomy, unreviewed escalation, or dignity-undermining workflow can emerge. In this sense, NAI 2.0™ is presented not as a self-sufficient technology, but as a constrained sociotechnical arrangement intended to support accountable eldercare practice.

## 5.2 System Boundaries

A defining characteristic of the proposed architecture is the explicit limitation of AI authority. NAI 2.0™ is designed as a bounded non-agentic system whose outputs remain advisory, structured, and review-dependent rather than self-executing. This distinction is foundational. The system may identify patterns, classify predefined risks, generate alerts, prioritize cases, or produce structured recommendations, but it does not independently authorize interventions, alter care plans, trigger irreversible treatment actions, or replace clinical or caregiving judgment in high-impact decisions.

The setting of system boundaries serves several functions. At the most basic level, it clarifies intended use. A system that attempts to do everything often becomes difficult to govern, difficult to validate, and difficult to constrain. By contrast, NAI 2.0™ is intentionally designed around a narrow and disciplined role within eldercare workflows. Its primary function is to support attention, structured review, and bounded decision assistance in situations where an older person's welfare may be affected by delay, omission, inconsistency, or fragmented information. It is not designed to become an autonomous care authority.

Within these boundaries, the system may perform four broad categories of activity. First, it may receive and organize predefined data inputs relevant to a care event, such as selected observational indicators, structured clinical parameters, risk flags, or operational notes. Second, it may process these inputs through bounded inference routines to produce classifications, prompts, or alerts. Third, it may route outputs through constitutional constraints that assess impact level and determine whether escalation, review, or pause is required. Fourth, it may record workflow events, authorizations, and interruptions in a traceable audit structure.

Equally important, the architecture is defined by what it excludes. The system does not independently initiate medication changes, force escalation decisions, make consent-sensitive determinations, restrict a person's choices, or trigger coercive interventions without accountable human action. It does not operate as a closed loop in which sensor input results automatically in consequential output. Nor does it collapse review into symbolic approval by placing humans at the end of a largely predetermined workflow. The system's role remains subordinate throughout.

This boundary-setting is particularly important in eldercare because high-impact decisions are rarely reducible to technical signal alone. An alert that appears clinically relevant may take on a different meaning when interpreted in light of family presence, communication barriers, distress, confusion, cultural expectations, environmental conditions, or the person's known preferences. A bounded system therefore makes room for contextual judgment rather than assuming that inferential confidence is equivalent to practical legitimacy.

The architecture also distinguishes between low-impact,

moderate-impact, and high-impact outputs. Low-impact outputs may include workflow prompts or informational summarization that does not materially alter care without further human engagement. Moderate-impact outputs may include recommendations that warrant timely review but do not immediately threaten safety or rights. High-impact outputs include any recommendation, classification, or escalation pathway that could materially affect health status, bodily integrity, privacy, dignity, autonomy, or access to care. These outputs trigger stronger controls, including pause, authorization, and interruption requirements.

By establishing explicit operational boundaries, the framework seeks to reduce ambiguity about where AI assistance ends and human authority begins. This is essential not only for safety, but also for accountability. A system can only be governed meaningfully if its powers are specified in advance. In this respect, system boundaries are not a peripheral technical detail; they are the first line of constitutional control.

## 5.3 Architectural Layers

The architecture of NAI 2.0™ is organized into a set of interdependent layers, each with a distinct function in maintaining bounded authority and meaningful human oversight. These layers should not be understood as isolated technical modules alone. They operate together as a governance structure through which data are restricted, inferences are constrained, escalation is disciplined, and accountability is preserved.

### 5.3.1 Data Input Layer

The data input layer governs what information enters the system and under what conditions. In line with the framework's minimum-necessary approach, this layer is intentionally restrictive. It accepts only predefined categories of data judged relevant to the specific eldercare use case. These may include selected physiological indicators, structured care observations, timing markers, clinician-entered notes in constrained fields, environmental flags, or workflow status data. The purpose is to avoid indiscriminate ingestion of data merely because such data are available.

This layer performs an important governance role because the scope of data intake shapes the scope of system influence. If an AI system absorbs excessive, poorly governed, or weakly contextualized information, it becomes harder to justify how outputs are produced and harder to control privacy exposure. By limiting inputs, the architecture reduces both technical noise and governance risk. It also supports clearer explanation of intended use, since the system's operational domain remains closely tied to defined data categories.

The data input layer may also include validation checks, timestamping, source verification, and completeness indicators. These functions do not guarantee data quality, but they improve the traceability of what the system received and what it did not receive. This is important in eldercare, where incomplete records, delayed observations, or inconsistent documentation can materially affect downstream interpretation.

### 5.3.2 Bounded Inference Layer

The bounded inference layer is where the system processes inputs and generates outputs such as classifications, prompts, or risk-oriented recommendations. The defining feature of this layer is not merely its analytical function, but the limits placed upon it. Its outputs are bounded in scope, expressed within predefined categories, and prevented from directly converting into autonomous action.

The architecture assumes that even a highly capable inference system should remain subordinate to the governance model within which it operates. Accordingly, the inference layer is constrained in at least three ways. First, it produces outputs only within the intended use domain defined by system boundaries. Second, it communicates outputs in a structured format that supports review rather than obscure probabilistic authority. Third, its outputs are not action-complete; they are always subject to constitutional checks and, where necessary, human interpretation.

This boundedness is significant because many risks associated with AI in care settings do not arise from inference alone, but from how inference is operationalized. A recommendation can become practically coercive if it is presented as inevitable, if the workflow is designed for rapid acceptance, or if users are not given room to challenge its basis. By restricting the inference layer to a non-executive role, the architecture attempts to preserve the distinction between analytic assistance and decision authority.

### 5.3.3 Constitutional Constraint Layer

The constitutional constraint layer is the normative core of the architecture. Its purpose is to assess AI outputs against predefined governance rules before they can influence workflow in consequential ways. This layer determines whether an output falls within low-impact operation, requires review, triggers Sacred Pause™, or becomes eligible for interruption or override through Sovereign Brake.

In practical terms, the constitutional layer functions as a rule-governed mediator between inference and action. It evaluates impact level, uncertainty, sensitivity, and escalation significance using predefined thresholds. These thresholds need not be fully automated in every implementation; what matters is that they are explicit, documented, and enforceable. This layer thereby transforms governance from a downstream human expectation into an upstream structural condition.

The constitutional constraint layer is especially important because it reduces the risk of silent authority expansion. Without such a layer, the logic of workflow efficiency may gradually shift AI outputs toward default acceptance. The constitutional model resists that drift by requiring the system itself to recognize categories of output that must not proceed casually. In this sense, it serves as the operational embodiment of the framework's non-agentic philosophy.

### 5.3.4 Hardware Enforcement Layer

The hardware enforcement layer distinguishes NAI 2.0™ from architectures that rely solely on software prompts or institutional policy. This layer is designed to impose stronger interruptibility and authorization conditions through physical, device-level, or tightly coupled infrastructural mechanisms. Its function is not to replace software safeguards, but to reinforce them where the consequences of bypass or drift are greatest.

The rationale for this layer is grounded in practical governance. In high-pressure care environments, software warnings may be ignored, poorly configured, or normalized into routine dismissal. Hardware-linked constraints can provide stronger assurance that certain pathways remain genuinely non-bypassable without human intervention. Depending on implementation, this may involve physical authorization devices, dedicated review terminals, secure activation controls, or other infrastructural mechanisms that require explicit human engagement before high-impact progression occurs.

This layer supports the framework's emphasis on sovereignty by ensuring that oversight is not merely symbolic. A human should not only be nominally responsible; the architecture should require their materially effective participation. Hardware enforcement helps create that condition.

### 5.3.5 Human Review Interface Layer

The human review interface layer is where authorized clinicians, caregivers, or designated oversight personnel engage with system outputs. In many AI deployments, interface design is treated as a usability concern. In this framework, it is also a governance concern. If the interface encourages speed over reflection, obscures uncertainty, or frames acceptance as default, then meaningful human oversight is weakened regardless of formal policy.

For this reason, the review interface should present outputs in a way that supports contextual understanding and active judgment. This may include clear indication of impact level, explanation of why a prompt was generated, visibility of uncertainty or confidence limits where appropriate, identification of required review steps, and accessible options to pause, escalate, reject, or request further input. The interface should not assume that efficiency alone is the highest value.

The role of this layer is therefore interpretive as well as procedural. It allows the human reviewer to bring contextual knowledge to bear on machine-generated output and to make decisions that reflect the lived realities of the person receiving care. In eldercare, that contextual role is indispensable.

### 5.3.6 Logging and Audit Layer

The logging and audit layer records the movement of information and decisions across the system. It captures relevant events such as data intake, inference generation, constitutional classification, pause triggers, authorization events, override activation, escalation steps, and final workflow outcomes. The purpose of logging is not merely retrospective documentation. It is a core mechanism of accountability, governance review, and continuous improvement.

In safety-critical settings, the absence of traceability makes it difficult to identify whether harm arose from poor input quality, flawed recommendation logic, rushed human approval, bypass of controls, or organizational misuse. The logging layer helps preserve this visibility. It also supports regulatory expectations relating to documentation, monitoring, and post-deployment governance.

Importantly, logging must itself be bounded and governed. Auditability should not become an excuse for excessive data retention or unnecessary privacy intrusion. Accordingly, the architecture assumes role-based access to audit records, proportionate retention policies, and alignment with applicable data governance requirements.

### 5.3.7 Escalation and Fallback Layer

The final layer is the escalation and fallback layer, which determines what happens when the system identifies a potentially consequential issue, encounters uncertainty, or is interrupted. A bounded architecture must not only prevent unsafe progression; it must also provide safe alternatives when progression is halted. This layer therefore routes high-impact outputs toward appropriate human review pathways and ensures that system uncertainty or failure does not produce operational paralysis.

Fallback mechanisms are especially important in eldercare because a system that pauses too readily without clear alternatives may burden staff, delay response, or create new forms of risk. Conversely, a system that never pauses may privilege speed over judgment. The escalation and fallback layer seeks to balance these concerns by ensuring that interruption leads into accountable human processes rather than into confusion or abandonment.

Taken together, these seven layers form an architecture in which AI remains analytically useful but governance remains structurally primary. The design goal is not friction for its own sake. It is disciplined friction at the points where care decisions become ethically, clinically, or dignity-relevant.

## 5.4 Care Event Workflow

To clarify how the architecture operates in practice, this section traces the workflow of a representative care event. The purpose of the example is not to describe a single universal

scenario, but to illustrate how bounded AI support, constitutional checks, and human authority interact in a typical high-impact eldercare pathway.

A care event begins when relevant information enters the system through the data input layer. This may occur through structured observation, routine documentation, predefined sensor input, or clinician-entered indicators associated with a change in an older person's condition or care context. At this stage, the system does not act. It receives and validates input within the categories defined by intended use.

The bounded inference layer then processes the input and generates a structured output. This output may take the form of a risk prompt, recommendation for further review, classification of concern, or signal that escalation criteria may be met. Crucially, the output is not itself an intervention. It is a bounded recommendation produced for further governance processing.

The output then passes to the constitutional constraint layer. Here, the system evaluates whether the output is low-impact, moderate-impact, or high-impact according to predefined criteria. If the output is low-impact, it may be routed to standard human review within ordinary workflow, accompanied by logging and traceability requirements. If the output is moderate-impact, the system may require acknowledgment by a designated reviewer before workflow continues. If the output is high-impact, Sacred Pause™ is triggered.

Once Sacred Pause™ is activated, the workflow cannot progress automatically. The case is held for explicit human review through the designated interface. The reviewer is presented with the relevant prompt, the basis for the system's concern within the limits of explainability appropriate to the implementation, and the available options. These options include acceptance of further review steps, rejection of the recommendation, escalation to senior oversight, or invocation of Sovereign Brake if safety, context, or consent concerns render the recommendation inappropriate.

If the reviewing human actor determines that the recommendation is reasonable and aligned with the individual's care context, progression may continue through a controlled authorization process. Where required,

Tiger .1x Key™ functions as the bounded authorization mechanism confirming that a qualified human has explicitly permitted the next step. That event is logged as an accountable transition rather than an assumed workflow continuation.

If, however, the reviewer judges that the recommendation is contextually unsound, dignity-undermining, premature, or otherwise unsafe, Sovereign Brake may be activated. This halts progression and routes the case into an alternative human-led process. In this way, the architecture ensures that machine output never becomes self-fulfilling merely because it has entered the workflow.

At the end of the event, the logging and audit layer records relevant workflow details, including input timing, output category, pause triggers, authorization status, interruption

events, and ultimate human action. These records support later review, quality improvement, and governance analysis.

What this workflow demonstrates is that the architecture is not designed for seamless automation. It is designed for controlled assistive operation. Its value lies in ensuring that consequential action remains reviewable, interruptible, and attributable to accountable human decision-makers rather than quietly absorbed into machine-led process.

## 5.5 Hardware-Enforced Safety Logic

A central innovation of the proposed framework is its reliance on

hardware-enforced safety logic in addition to software safeguards and procedural oversight. The underlying premise is that in high-impact eldercare environments, policy alone may be insufficient to preserve meaningful human control. Staff work under time pressure, organizations face efficiency demands, and routine practices can normalize the rapid acceptance of machine output. Under such conditions, purely software-based warnings or interface notices may be ignored, bypassed, or gradually treated as administrative friction rather than serious governance controls.

The architecture responds to this problem by introducing stronger forms of interruptibility and authorization that are linked to physical or infrastructural constraints. The precise technical form of these constraints may vary by deployment context, but the principle remains consistent: certain transitions must not be executable unless an accountable human actor performs an explicit and materially effective authorization step. High-impact outputs should not move from inference to action by default, nor should interruptive safeguards be easy to disable through convenience-driven workflow adaptation.

This hardware-linked model matters for several reasons. First, it reduces the likelihood of silent override. A software safeguard can often be dismissed with a click; a hardware-linked authorization requirement creates a more deliberate transition. Second, it improves the enforceability of pause conditions. If Sacred Pause™ can be ignored without friction, then its role becomes symbolic. Hardware reinforcement helps ensure that a pause is operationally real. Third, it strengthens accountability by making authorization events more discrete, traceable, and role-specific.

The framework does not assume that hardware enforcement is infallible. Poor implementation, inadequate maintenance, or inappropriate institutional use could still undermine its value. Moreover, excessive rigidity can produce burden if not carefully aligned with workflow realities. The argument is therefore not that hardware enforcement automatically solves the governance problem, but that it offers a stronger layer of protection where bypassable software controls may be insufficient.

Hardware-enforced safety logic is particularly significant in eldercare because many high-impact situations involve ambiguity rather than obvious emergency. In such cases, the greatest risk may not be malicious misuse but routine acceleration. A recommendation

that should trigger reflection may instead become normalized as an expected next step. Hardware-linked interruption can slow this transition sufficiently to restore practical judgment. In this sense, the architecture values deliberate review over frictionless execution.

The design also supports the principle of bounded activation. AI capabilities may be available within the system, but not all capabilities are continuously or equally actionable. Some outputs require additional human validation not only because they are technically uncertain, but because their consequences are morally or clinically significant. By linking activation to explicit authorization, the system preserves the distinction between what can be computed and what may legitimately proceed.

This logic also reinforces the concept of sovereignty introduced in Chapter 3. A human is sovereign over a system not merely because policy says so, but because the architecture makes certain forms of action impossible without that person's involvement. Hardware-enforced safety therefore serves as a practical expression of constitutional governance: it embeds normative limits in the operating conditions of the system itself.

Ultimately, the role of hardware-enforced logic in NAI 2.0™ is to support a safer and more disciplined relationship between inference and action. It does not eliminate error, nor does it replace organizational responsibility. What it does is reduce the probability that high-impact AI outputs will bypass reflective human judgment through speed, habit, or design weakness. In eldercare, where the consequences of inappropriate progression may include both physical harm and dignity erosion, that reduction is ethically and operationally significant.

## 5.6 Privacy and Data Handling Architecture

The privacy and data handling architecture of NAI 2.0™ is guided by the principle that eldercare AI should use only the data necessary to support its bounded function and should do so within a framework of access control, segmentation, traceability, and proportionality. This reflects both regulatory concerns and the ethical reality that older adults may be especially vulnerable to the harms of unnecessary surveillance, opaque data use, or over-collection of sensitive information.

At the center of this architecture is a minimum-necessary approach. Data intake is limited to categories directly relevant to the system's intended use, and collection should not expand merely because technical capability permits broader capture. This is important not only for privacy protection, but also for governance clarity. A system with uncontrolled data appetite is harder to justify, harder to explain, and more prone to mission drift.

The architecture also assumes segmentation of access. Not every user should see every class of information, and not every system component should process all available data. Role-based access control therefore forms a core part of the deployment model. Clinicians, caregivers, reviewers, and administrators may require different levels of

visibility depending on their operational role. Such differentiation supports confidentiality while also reducing the risk of casual overexposure to sensitive information.

Where feasible, the framework favors localized or tightly governed processing pathways over indiscriminate data transfer across loosely controlled environments. This does not require that all implementations be fully local in every case, but it does require that data movement be justified, documented, and minimized. The architecture is therefore compatible with privacy-by-design principles insofar as it seeks to reduce unnecessary transfer, limit retention, and align access with explicit governance purposes.

Logging is also part of privacy governance. Audit trails should record not only workflow decisions but also access events, authorization actions, and significant system interactions. However, auditability must remain proportionate. The framework does not endorse unlimited retention or unrestricted internal visibility. Rather, it assumes governed logging, appropriate retention schedules, and controlled access to records based on legitimate oversight needs.

An additional concern is the relationship between privacy and dignity. In eldercare, data governance is not merely a matter of compliance. Excessive monitoring, poorly explained data capture, or unnecessary visibility into intimate aspects of life may produce dignity-related harms even where security is technically strong. For this reason, the architecture treats privacy as a relational as well as technical concern. Data handling should support care without normalizing unwarranted intrusion.

The chapter does not claim that privacy can be guaranteed by architecture alone. Lawful basis, organizational governance, consent management where applicable, and jurisdiction-specific requirements remain essential. Nonetheless, the proposed design seeks to create a compliance-supportive environment by ensuring that data use is constrained, access is controlled, actions are traceable, and the system's informational footprint remains proportionate to its bounded role.

## 5.7 Failure Modes and Risk Controls

No governance architecture is risk-free, and it would weaken the thesis to suggest otherwise. The appropriate question is not whether NAI 2.0™ eliminates risk, but whether it changes the structure of risk in a way that reduces preventable harm and improves accountability. Several failure modes remain possible even within a bounded non-agentic architecture.

One risk is inappropriate recommendation generation. The system may produce an alert or recommendation that is technically plausible but contextually wrong due to incomplete data, misclassification, limited generalizability, or subtle care factors not captured within the model. This risk is addressed partly through bounded authority: recommendations do not self-execute. It is also addressed through human review, pause mechanisms, and the capacity to reject or interrupt workflow.

A second risk is alert fatigue or interruption fatigue. If the system produces too many prompts or triggers pause mechanisms too frequently, human reviewers may become desensitized or frustrated. This can undermine governance by encouraging mechanical approval or informal workarounds. The architecture seeks to mitigate this through impact-based thresholding, role-sensitive escalation design, and careful calibration of what counts as a high-impact trigger. Even so, this remains a real implementation challenge.

A third risk is human over-reliance on AI output. Even where humans remain formally in control, they may defer too readily to machine-generated recommendations, especially under workload pressure. The framework responds through interface design that supports reflection, through explicit pause conditions, and through structurally meaningful override tools such as Sovereign Brake. Yet over-reliance cannot be fully eliminated by architecture alone; it also requires training and institutional culture.

A fourth risk is delayed action caused by excessive caution. A bounded system that requires pause and authorization may improve oversight but could also slow response in cases where timely action is needed. This tension is especially important in healthcare settings. The proposed model therefore depends on careful definition of threshold categories, so that interruption is reserved for genuinely high-impact or uncertain cases rather than routine workflow. The architecture aims for disciplined caution, not blanket delay.

A fifth risk concerns dignity erosion through rigid workflow. Ironically, a system designed to preserve dignity could undermine it if its procedures become too inflexible, too bureaucratic, or too detached from lived care relationships. For example, excessive emphasis on formal review could burden staff in ways that reduce direct human presence. This risk underscores the need to treat architecture as part of a sociotechnical arrangement rather than as a complete solution.

Finally, there is the risk of implementation drift. Over time, local practices may weaken controls, reinterpret thresholds, or normalize shortcuts. Logging, auditability, and hardware-enforced safeguards help resist this drift, but they do not abolish it. Ongoing governance review remains necessary.

The broader point is that risk control in NAI 2.0™ operates through bounded authority, explicit interruption, accountable authorization, traceable workflow, and governance-aware design. These measures do not remove uncertainty, but they aim to ensure that uncertainty does not become invisible or operationally unchecked.

## 5.8 Practical Deployment Assumptions

The practical deployment of NAI 2.0™ depends on organizational conditions as much as on technical design. A bounded architecture cannot function as intended if deployed into an environment that lacks training, governance culture, maintenance capacity, or clearly

assigned responsibility. For this reason, the framework assumes several practical preconditions.

First, deployment requires staff training that goes beyond basic interface use. Users must understand the purpose of bounded operation, the meaning of pause and override functions, the distinction between recommendation and authority, and the importance of preserving contextual judgment in eldercare. Without such understanding, even well-designed safeguards may be reduced to procedural inconvenience.

Second, the organization must maintain clear governance protocols defining roles, escalation routes, authorization rights, and accountability structures. A system built around meaningful human oversight cannot operate effectively if no one is certain who has the authority to approve, interrupt, or review high-impact outputs.

Third, implementation depends on infrastructural reliability and maintenance. Hardware-enforced features, logging systems, access controls, and review interfaces all require ongoing support. If these degrade, the framework's governance claims weaken correspondingly.

Fourth, the model assumes an institutional culture in which

interruption is treated as legitimate rather than as inefficiency. In some care environments, speed is rewarded so strongly that any friction is viewed negatively. NAI 2.0™ requires a different orientation: one in which deliberate pause at consequential points is recognized as a feature of safe and dignified practice.

Finally, the architecture presumes a commitment to continuous review. Thresholds, workflows, and authorization patterns should be revisited in light of observed operation, audit findings, and stakeholder feedback. The framework is therefore not static. Its boundedness must be preserved actively through governance as well as design.

**Concluding Note to Chapter 5**

This chapter has translated the conceptual claims of the thesis into a concrete system architecture and deployment model. It has argued that eldercare AI should be designed not as a self-executing authority, but as a bounded and review-dependent component of a wider sociotechnical care environment. To that end, the chapter has defined explicit system boundaries, described the architecture's seven functional layers, illustrated the workflow of a representative care event, and explained the role of hardware-enforced safety logic in preserving meaningful human sovereignty.

The chapter has also shown that technical design cannot be separated from governance. Data restriction, constitutional checks, authorization pathways, interruptibility, auditability, and fallback procedures are all architectural expressions of the broader argument that safety and dignity must be embedded into operation rather than added rhetorically after the fact. At the same time, the discussion of failure modes and deployment assumptions

has made clear that architecture alone is insufficient. Training, institutional culture, maintenance, and ongoing oversight remain indispensable.

Taken together, the chapter establishes the technical credibility of NAI 2.0™ as a governance-oriented eldercare architecture. It demonstrates that non-agentic design can be specified in operational terms rather than left at the level of principle. With this architectural foundation in place, the next chapter can examine how the framework aligns with regulatory, ethical, and governance expectations across healthcare AI deployment contexts.

# Chapter 6. Regulatory, Ethical, and Governance Alignment

## 6.1 Introduction

This chapter evaluates NAI 2.0™ against the regulatory, ethical, and governance expectations most relevant to AI deployment in eldercare. Its purpose is not to claim that the framework is automatically compliant by virtue of design alone, nor to suggest that technical architecture can substitute for legal review, organizational accountability, or formal certification. Rather, the chapter examines whether the proposed framework has been constructed in a manner that is compliance-supportive, governance-ready, and ethically defensible within the context of healthcare AI deployment.

The need for this analysis arises from a recurring weakness in contemporary AI governance discourse. Many systems are described as safe, ethical, or trustworthy in broad terms, yet the relationship between technical design and concrete regulatory obligations often remains underdeveloped. Conversely, regulatory discussions may specify abstract requirements for risk management, human oversight, logging, transparency, or accountability without showing how those principles can be operationalized in real system architecture. This thesis seeks to narrow that gap. As argued in earlier chapters, eldercare is not simply another domain of digital deployment. It is a context involving vulnerability, bodily welfare, privacy sensitivity, fluctuating capacity, and dignity-related concerns that make governance architecture especially important.

The chapter proceeds by examining the framework through several lenses. It begins with the principle of governance-by-design, showing how NAI 2.0™ embeds oversight, bounded authority, and interruptibility into its technical structure. It then considers the framework across the AI and medical software lifecycle, including design, validation, deployment, monitoring, and change management. After that, it discusses the relevance of a risk-based Software as a Medical Device logic, with particular attention to HSA Class B-style regulatory reasoning as a useful comparator for moderate-risk clinical support systems. The chapter then turns to broader AI governance, especially the EU AI Act, examining how the proposed architecture aligns with expectations relating to high-risk systems, human oversight, traceability, and risk control. It proceeds to data governance by assessing how the framework supports privacy and accountability principles associated with

HIPAA, GDPR, and CCPA. Finally, the chapter evaluates the model through an ethical governance lens, focusing on dignity, autonomy, accountability, privacy, and non-maleficence before acknowledging unresolved tensions and residual gaps.

The central argument advanced throughout is that NAI 2.0™ is best understood as a governance architecture rather than merely a technical system. Its strongest alignment claim is not that it already satisfies every jurisdiction-specific requirement in operational detail, but that its design is intentionally structured to support the kinds of control, oversight, boundedness, traceability, and restraint increasingly demanded in healthcare AI

regulation. In this sense, the chapter contributes to the thesis by showing that the framework is not only conceptually coherent and technically bounded, but also meaningfully legible to emerging regimes of healthcare and AI governance.

## 6.2 Governance-by-Design

A central premise of this thesis is that governance should be embedded into system architecture rather than appended as a downstream procedural layer. This principle, referred to here as governance-by-design, holds that ethical and regulatory expectations are more likely to be operationally effective when they are incorporated into the structure of the system itself. In eldercare, this is particularly important because many of the most serious risks do not arise from explicit malicious use but from routine drift, workflow pressure, automation bias, poorly timed escalation, or the gradual normalization of machine-led decision pathways.

Within NAI 2.0™, governance-by-design is expressed through several architectural choices. First, the system is intentionally

non-agentic. This means that AI outputs remain bounded and advisory rather than self-executing. The significance of this design choice is regulatory as well as ethical: it narrows the intended use of the system, limits the scope of autonomous authority, and preserves a clearer distinction between machine assistance and human decision-making. In governance terms, this reduces ambiguity about responsibility and supports a more proportionate control model.

Second, the framework incorporates constitutional constraints that classify outputs by impact and determine whether additional review, interruption, or authorization is required. This reflects a shift away from governance understood solely as policy or training. Instead, the architecture itself recognizes that not all outputs are equal and that some require stronger procedural friction. This is particularly relevant in eldercare, where high-impact outputs may implicate safety, privacy, autonomy, or dignity in ways that cannot be responsibly managed through ordinary workflow alone.

Third, governance-by-design is embodied in the system's

interruptibility mechanisms, particularly Sacred Pause™ and

Sovereign Brake. These mechanisms are not optional user features but structural protections intended to ensure that consequential AI outputs remain subject to meaningful human control. In regulatory terms, they support human oversight. In ethical terms, they preserve opportunities for contextual judgment, refusal, deferral, and explanation. In governance terms, they make it harder for machine output to become practically authoritative through speed or habit.

Fourth, the architecture includes bounded authorization through

Tiger .1x Key™, as well as logging and audit functions that preserve traceability of key workflow transitions. These features support accountability by linking consequential

system progression to identifiable human action. They also support later investigation, quality review, and post-deployment governance.

The importance of governance-by-design lies in the fact that healthcare AI systems do not operate in neutral environments. Eldercare settings may involve high workload, staffing constraints, emotionally charged decisions, and asymmetries of power between institutions and older adults. Under such conditions, ethical intention alone may be insufficient. A framework that requires oversight structurally is more likely to preserve oversight in practice than one that merely advises users to behave carefully.

At the same time, governance-by-design does not eliminate the need for external governance. Legal interpretation, institutional policy, ethics review, procurement standards, user training, clinical leadership, and organizational culture remain essential. The argument is therefore not that design replaces governance, but that it makes governance more credible by embedding it into operational pathways. NAI 2.0™ is thus presented as a framework in which governance is not an afterthought. It is part of the system's constitutional structure.

## 6.3 Lifecycle Alignment

An important measure of governance maturity in healthcare AI is whether a system can be evaluated across its entire lifecycle rather than only at the point of initial design. For this reason, NAI 2.0™ is considered here against a lifecycle alignment model spanning design, development, validation, deployment, monitoring, and change management. This approach reflects the broader regulatory expectation that safety and accountability are not static properties. They must be preserved over time as systems are implemented, used, updated, and interpreted within real care environments.

**Design Stage**

At the design stage, the framework aligns with governance expectations by defining a narrow intended use, explicit system boundaries, and a non-agentic operational role. These choices are important because regulatory and ethical control depend heavily on clarity of purpose. A system with unclear scope is difficult to classify, difficult to test, and difficult to govern. By contrast, NAI 2.0™ specifies from the outset that it is a bounded support architecture for eldercare rather than a self-governing care engine.

This stage also includes the embedding of core risk controls into architecture. Constitutional constraints, pause mechanisms, override pathways, and authorization requirements are not retrofitted after deployment; they are part of the system's design logic. This supports the principle that safety and human oversight should be planned rather than improvised.

**Development Stage**

At the development stage, lifecycle alignment requires that system implementation remain faithful to intended use and documented governance assumptions. In the case of NAI 2.0™, this means that data pathways, model outputs, interface functions, and escalation thresholds must be built in a way that preserves boundedness. Development should not widen system authority informally or permit hidden routes through which advisory outputs become executive decisions.

This stage also requires documentation. Even where the thesis does not claim completion of a regulated product development process, the architecture clearly implies the need for requirements specification, design records, risk control mapping, and role-based access definitions. From a governance standpoint, development should produce not only a functioning system but also an intelligible record of how the system is intended to behave.

**Validation Stage**

At the validation stage, the framework aligns with governance expectations by supporting scenario-based testing, threshold review, workflow simulation, and examination of interruption and authorization pathways. Validation in this thesis is not limited to model accuracy. It also includes testing whether human oversight is operationally meaningful, whether high-impact outputs trigger pause conditions as intended, and whether the system behaves in a manner consistent with its non-agentic claims.

This broader notion of validation is especially important in eldercare. A system could perform well in narrow analytic terms yet still fail governance validation if it pushes staff toward unreflective reliance or if it undermines dignity-sensitive care processes. NAI 2.0™ therefore supports validation as a sociotechnical exercise rather than merely a technical benchmark.

**Deployment Stage**

At the deployment stage, lifecycle alignment is reflected in the framework's emphasis on training, role clarity, infrastructure reliability, and bounded access. Deployment is where governance promises are most often tested. A system that looks safe in principle may become unsafe in practice if users are not properly trained, thresholds are misunderstood, or interruptive mechanisms are treated as workflow obstacles rather than legitimate safeguards.

The architecture responds by requiring explicit human review points, role-sensitive authorization, and traceability of consequential steps. This supports accountability during live use and helps preserve the system's constitutional character even under organizational pressure.

**Monitoring Stage**

At the monitoring stage, the framework aligns with governance expectations through logging, auditability, and post-deployment review. Monitoring is necessary because the behavior of a system in live practice may diverge from expectations formed during design and validation. Staff may develop workarounds, alert patterns may change, or certain controls may prove too weak or too burdensome. The logging and audit structure of NAI 2.0™ is intended to make such developments visible.

Monitoring also supports reflective governance. By examining pause frequency, authorization patterns, override use, and workflow outcomes, an organization can assess whether the system is preserving meaningful oversight or drifting toward formal but ineffective control.

**Change Management Stage**

Finally, at the change management stage, lifecycle alignment requires that modifications to thresholds, interfaces, model behavior, data sources, or workflow integration be treated as governance-relevant events rather than routine technical adjustments. In healthcare AI, small changes can have significant consequences if they affect output interpretation, review burden, or the conditions under which action occurs.

The architecture of NAI 2.0™ supports change-sensitive governance because it depends on explicit rules, thresholds, and bounded authority structures. Any material change that alters these features should trigger renewed review, documentation, and, where appropriate, re-validation. This is consistent with the broader regulatory principle that safety cannot be assumed to persist unchanged across evolving system states.

Taken together, this lifecycle analysis suggests that NAI 2.0™ aligns well with the growing expectation that healthcare AI should be governable from conception through post-deployment oversight. Its architecture is particularly strong in relation to intended use clarity, human oversight, traceability, and change-sensitive governance. Its weaker point, at least at the thesis stage, is that lifecycle alignment is presented analytically rather than demonstrated through full-scale commercial deployment evidence. That limitation should be acknowledged, but it does not diminish the framework's value as a governance-oriented design model.

## 6.4 HSA Class B SaMD Relevance

A useful way to assess the regulatory maturity of NAI 2.0™ is to consider its relevance to a risk-based Software as a Medical Device logic, particularly the reasoning associated with HSA Class B-type systems. The aim of this section is not to classify the framework definitively, since formal classification depends on intended use, jurisdiction, implementation details, and regulatory engagement. Rather, the objective is to examine whether the architecture has features consistent with the kind of control environment

expected of software that may inform clinical or care-related decisions without independently assuming full therapeutic authority.

The relevance of a Class B-style SaMD perspective lies in the fact that NAI 2.0™ is designed for healthcare-adjacent and care-informing contexts where outputs may influence decisions affecting older adults. Even if the system remains non-agentic, its recommendations or alerts may still be consequential. Regulatory reasoning in this space typically turns on factors such as intended use, significance of the information provided, severity of the condition or situation addressed, extent of human oversight, and the potential consequences of misuse or error.

**Intended Use and Scope**

From a SaMD perspective, intended use is foundational. NAI 2.0™ is intentionally designed as a bounded decision-support and governance framework rather than an autonomous diagnostic or treatment system. This narrow scope is important because it limits claims and constrains operational authority. The framework does not purport to replace clinicians, prescribe treatment independently, or function as a closed-loop intervention engine. Such boundedness is likely to support a more proportionate regulatory profile than would be the case for a system claiming direct therapeutic autonomy.

At the same time, because the system may influence care pathways, risk attention, or escalation review, it still falls within a domain where governance must be rigorous. A system does not need to act independently to create clinically relevant consequences. This is why the architecture's emphasis on review, documentation, and interruption remains regulatory significant.

**Risk Control and Human Oversight**

A core strength of the framework, viewed through SaMD logic, is its commitment to meaningful human oversight. Many regulatory systems place importance on whether software outputs are reviewed by qualified humans before consequential action occurs. NAI 2.0™ aligns strongly with this expectation because high-impact outputs are structurally prevented from moving forward without human engagement. Sacred Pause™, Sovereign Brake, and Tiger .1x Key™ collectively provide a more robust oversight environment than a conventional alert-and-acknowledge model.

This matters because in medical software governance, oversight must be practical rather than fictional. If a human is technically responsible but not structurally empowered to interrupt or interpret outputs, the claim of review becomes weak. By embedding oversight in architecture, the framework supports a more credible human-governed model.

**Documentation and Traceability**

Regulated software environments also require documentation sufficient to explain intended use, risk controls, system behavior, and accountability pathways. While this thesis does not claim full product documentation as would be required in a commercial approval setting, the architecture clearly lends itself to such documentation. The framework defines distinct layers, explicit trigger logic, role-based authorization points, and logging requirements. This provides a strong basis for a future technical file, risk management dossier, or safety case.

Traceability is particularly strong within the design. Input events, output categories, pause triggers, override actions, and authorization transitions can all be recorded and audited. This supports both premarket accountability and post-deployment review.

**Quality Management Implications**

A further area of relevance concerns quality management. Any real-world implementation of NAI 2.0™ in a regulated healthcare context would need to sit within a quality system capable of governing design controls, testing, release management, incident review, and corrective action. The architecture is compatible with this requirement because it treats change, escalation, and authorization as governed processes rather than informal technical settings.

However, it must be emphasized that compatibility with a quality management approach is not the same as evidence that such a system has been fully implemented. The framework as presented in the thesis supports quality-oriented governance conceptually and structurally, but actual compliance would depend on organizational adoption and documented process discipline.

**Post-Deployment Monitoring**

SaMD logic increasingly emphasizes post-deployment monitoring, especially where software behavior may interact with evolving clinical practice. NAI 2.0™ supports this through its logging and audit layer, which can help organizations detect unexpected usage patterns, excessive pause frequency, underuse of override mechanisms, or drift in authorization behavior. These capabilities support a monitoring orientation consistent with regulated healthcare software expectations.

**Overall Relevance**

Overall, the framework shows substantial alignment with the governance logic one would expect from a moderate-risk clinical decision-support environment. Its strongest features are intended use clarity, bounded authority, enforced human oversight, and traceability. Its limitations are those of the thesis as a research artefact rather than a fully certified product: formal classification remains context-specific, product testing remains partial, and regulatory approval would require institution-specific evidence and process controls beyond the scope of this dissertation.

For that reason, the most defensible conclusion is that NAI 2.0™ demonstrates SaMD-relevant design maturity rather than formal classification status. It is structured in a way that supports risk-sensitive healthcare governance and would be capable, in principle, of being developed within a regulated software framework, provided that future implementation work includes the necessary documentation, validation, and quality management processes.

## 6.5 EU AI Act Alignment

The EU AI Act provides one of the most significant contemporary reference points for AI governance, particularly in relation to systems that may qualify as high-risk because of their use context, their likely effects on individuals, or their role in sensitive domains. Although this thesis does not claim a formal legal determination under the Act, the framework of NAI 2.0™ can be meaningfully assessed against several of the governance principles associated with high-risk AI. These include risk management, human oversight, transparency, logging, accountability, and attention to vulnerable individuals.

### High-Risk Context Logic

Eldercare is a domain in which the use of AI may reasonably attract heightened scrutiny because it involves individuals who may be vulnerable by reason of age, dependency, health status, or social situation. Where AI outputs can influence access to care, escalation, monitoring intensity, or interpretation of need, the governance burden increases. Even if NAI 2.0™ remains non-agentic, it operates in a context where harm could arise if outputs are misused, over-relied upon, or allowed to shape care without appropriate review.

The architecture responds to this high-risk logic by constraining authority from the outset. Rather than attempting to justify expanded autonomy with later safeguards, the framework assumes that vulnerability should trigger stronger boundedness. This is highly consistent with the spirit of the EU AI Act, which emphasizes risk mitigation and human control in sensitive applications.

### Risk Management

The Act's broader governance orientation encourages ongoing risk management across the lifecycle of AI systems. NAI 2.0™ aligns with this principle through its impact classification logic, failure-mode awareness, interruption pathways, and monitoring structure. Risk is not treated as a one-time technical assessment but as a continuing governance issue shaped by workflow, implementation, and changing practice.

This is particularly evident in the constitutional layer of the architecture. By classifying outputs and reserving stronger controls for higher-impact cases, the framework

operationalizes risk sensitivity rather than assuming uniform treatment of all outputs. This supports a proportionate but cautious design model.

### Human Oversight

Human oversight is among the most prominent governance requirements in contemporary AI regulation, and it is one of the areas where NAI 2.0™ aligns most strongly. The framework does not merely preserve a nominal human role. It uses Sacred Pause™, Sovereign Brake, and Tiger .1x Key™ to ensure that high-impact outputs remain reviewable, stoppable, and authorization-dependent.

This alignment is important because high-risk AI governance increasingly rejects the fiction that human presence alone is sufficient. Effective oversight requires timing, authority, and operational feasibility. NAI 2.0™ supports all three by interrupting consequential progression at the moment it matters and by preserving identifiable human control over whether workflow continues.

### Transparency and Traceability

The EU AI Act also places emphasis on transparency and record-keeping appropriate to system risk. NAI 2.0™ supports this through structured outputs, role-based review pathways, and a logging layer that records key workflow transitions. While transparency does not require complete technical exposition to every user, the architecture is designed to make the practical meaning of system outputs understandable enough for human review and accountability.

Traceability is especially well supported. The system can record when outputs were generated, how they were classified, whether pause conditions were triggered, who authorized progression, and whether an override occurred. This contributes to governance review and incident investigation in ways consistent with high-risk AI accountability principles.

### Accountability and Vulnerable Persons

A further area of alignment concerns accountability for systems affecting vulnerable individuals. The framework directly addresses this by treating eldercare not as a neutral deployment environment but as one requiring stronger safeguards due to dependency, dignity sensitivity, and the risk of coercive or opaque system influence. The architecture's emphasis on bounded authority and interruption is therefore not incidental; it is a governance response to vulnerability.

### Limits of Alignment

Despite these strengths, it would be misleading to claim full legal alignment in a definitive sense. The EU AI Act operates within a detailed legal framework that includes obligations extending beyond architecture alone, including provider responsibilities, conformity assessment, governance processes, documentation standards, and jurisdiction-specific interpretation. NAI 2.0™ aligns conceptually and structurally with many of the Act's core principles, but actual compliance would depend on the details of deployment, provider role, technical implementation, and evidentiary documentation.

The most accurate conclusion, therefore, is that NAI 2.0™ demonstrates

strong design-level consistency with the governance logic of high-risk AI regulation, especially in relation to human oversight, bounded authority, risk management, and traceability. This reinforces the thesis's broader argument that constitutional, non-agentic architecture can serve as a practical bridge between ethical aspiration and regulatory expectation.

## 6.6 HIPAA, GDPR, and CCPA Alignment

The data governance dimension of NAI 2.0™ is particularly important because eldercare AI necessarily engages with sensitive personal information, often including health-related data, behavioral observations, care status, and workflow records linked to identifiable individuals. In such settings, privacy cannot be treated as a secondary concern. It is both a legal matter and an ethical matter, especially where excessive monitoring or opaque data use may compromise dignity even in the absence of formal security breach. This section evaluates the framework against the broad governance principles associated with HIPAA, GDPR, and CCPA.

### 6.6.1 HIPAA

From a HIPAA-oriented perspective, the strongest alignment features of NAI 2.0™ are its commitment to access control, auditability, bounded data use, and a minimum-necessary approach. The architecture assumes that only specified categories of data relevant to the system's intended use should enter the system and that not all users should have the same level of visibility. This is consistent with a privacy and security model in which access is role-based and functionally justified rather than generalized.

The framework also supports audit logging, which is relevant to accountability and security monitoring. By recording key interactions, authorization steps, and workflow transitions, the system creates a traceable environment that can assist with review of inappropriate access, misuse, or irregular workflow behavior. Such traceability is important where health-related decision support may shape care processes affecting vulnerable individuals.

A further area of alignment lies in the architecture's segmentation logic. The system is designed so that data pathways and user roles can be differentiated rather than flattened

into one broad access environment. This supports confidentiality by limiting unnecessary exposure to sensitive information.

At the same time, HIPAA compliance in practice depends on organizational safeguards beyond architecture alone. Administrative controls, workforce training, breach response procedures, contractual arrangements, and infrastructure security all remain essential. NAI 2.0™ can be said to support HIPAA-style governance principles, particularly those related to minimum necessary access, security awareness, and accountability, but it does not by itself constitute proof of compliance in any specific deployment.

### 6.6.2 GDPR

The GDPR introduces a broader and in some respects more demanding model of data governance, especially where identifiable health-related information and vulnerable persons are involved. Several aspects of NAI 2.0™ align well with GDPR principles, particularly data minimization, purpose limitation, accountability, and restricted access.

The architecture's bounded intended use supports purpose limitation by clarifying why data are processed and restricting the system from expanding into loosely related functions without governance review. Its minimum-necessary data intake model supports data minimization, which is especially important in eldercare contexts where it may be technically possible to collect highly granular or intimate data but ethically unjustified to do so.

The framework's logging and role-based review structures also support accountability. Because the system records authorization points, pause triggers, and key workflow events, it is better positioned to support explanations of how decisions were influenced and who was responsible for consequential progression. This does not automatically satisfy all GDPR expectations relating to explainability or data subject rights, but it creates a more governable basis than opaque or unlogged systems.

Transparency under GDPR is also relevant. While NAI 2.0™ is not presented as a full user-facing communication policy, its architecture is compatible with a governance approach in which individuals, families, or lawful representatives can be informed that AI-supported processes are being used in a bounded decision-support role rather than as autonomous decision-makers. This matters because transparency is not only about technical documentation; it is also about ensuring that affected persons are not unknowingly subjected to consequential machine-led processes.

However, several GDPR matters remain deployment-specific. These include lawful basis for processing, handling of special category data, data subject rights management, retention policies, and international transfer arrangements where applicable. The framework supports these concerns structurally by limiting data flows and preserving

auditability, but actual GDPR compliance would depend on the legal and institutional context in which the system is used.

Accordingly, the most defensible claim is that NAI 2.0™ demonstrates

GDPR-supportive architectural principles, especially around minimization, bounded purpose, access control, and accountability. It should not be described as GDPR-compliant in the abstract absent deployment-specific governance arrangements.

### 6.6.3 CCPA

The CCPA provides a somewhat different but still important governance lens, especially in relation to transparency, user rights, and responsible data handling practices. NAI 2.0™ aligns with this general orientation in several ways.

First, the framework supports constrained data handling by limiting input categories and tying processing to explicit eldercare functions. This reduces the likelihood of expansive secondary use and supports clearer internal governance around what data are collected and why. Second, its logging and access control features support an environment in which disclosures about data use, internal handling, and accountability can be made more concretely than in systems lacking traceability.

Third, the architecture is compatible with the broader principle that individuals should not be subject to unbounded or opaque uses of personal information. Because NAI 2.0™ is designed around narrow purpose and bounded authority, it is less likely to drift into generalized profiling or unrelated secondary functions than more open-ended AI environments.

As with HIPAA and GDPR, however, actual CCPA compliance would depend on organizational implementation, disclosure practices, rights-response processes, and jurisdiction-specific operational details. The architecture supports good data governance practice, but it does not substitute for legal compliance activity.

**Overall Data Governance Position**

Across HIPAA, GDPR, and CCPA, a consistent pattern emerges. NAI 2.0™ is strongest where data governance intersects with boundedness,

role-sensitive access, traceability, and minimum-necessary use. These features align well with modern privacy expectations and are especially appropriate in eldercare, where privacy violations may become dignity violations as well.

The framework is weaker only in the sense that architecture alone cannot resolve all legal questions. Lawful basis, data subject rights, consent arrangements where relevant, retention schedules, third-party processing obligations, and cross-border issues all require

institution-specific governance. The thesis therefore does not claim universal data compliance. It argues instead that NAI 2.0™ creates a more privacy-conscious and accountability-supportive infrastructure than systems built around unrestricted data ingestion and opaque decision pathways.

## 6.7 Ethical Governance Analysis

Regulatory alignment, while essential, is not sufficient to establish the ethical adequacy of AI in eldercare. Law and ethics overlap, but they are not identical. A system may satisfy many procedural requirements while still undermining the quality of care relationships, eroding dignity, or normalizing subtle forms of coercion. For this reason, NAI 2.0™ must also be evaluated through a broader ethical governance lens, particularly in relation to autonomy, dignity,

non-maleficence, privacy, accountability, and the preservation of meaningful human presence.

### Autonomy and Meaningful Human Control

A major ethical concern in eldercare AI is whether technology preserves or diminishes the autonomy of those affected by its outputs. In the present framework, autonomy is protected indirectly and directly. It is protected indirectly because the architecture prevents AI from becoming a self-executing authority in matters that may shape care. It is protected directly because pause and review mechanisms preserve opportunities for explanation, contextual interpretation, and refusal.

This does not mean that the system guarantees autonomy in every case. Some older adults may have limited decision-making capacity, and care decisions may involve legally authorized representatives or professional obligations. Nevertheless, the architecture is ethically preferable to more agentic models because it does not collapse machine recommendation into de facto command. It preserves a space in which autonomy can still be recognized and negotiated.

### Dignity Preservation

The framework is especially strong in relation to dignity preservation, which the thesis has treated as a system-level concern rather than a purely interpersonal virtue. NAI 2.0™ supports dignity by resisting rushed, opaque, or machine-dominant progression in high-impact situations. Sacred Pause™ in particular can be understood as an ethical mechanism because it creates room for explanation, deliberation, and human recognition before action occurs.

Similarly, the non-agentic structure helps prevent the older person from being treated merely as the endpoint of optimization. By preserving human sovereignty and contextual

judgment, the architecture recognizes that dignified care often depends on relational understanding rather than computational efficiency alone.

### Non-Maleficence and Prevention of Harm

From the perspective of non-maleficence, the framework's core ethical value lies in its attempt to reduce preventable harm arising from over-automation, misplaced trust, or unreviewed escalation. Its bounded authority model, interruptibility, and traceability all support a cautious approach to harm reduction.

At the same time, the framework acknowledges that ethical harm can arise from both action and inaction. Excessive pause, over-triggering, or burdensome review procedures could delay response or create workflow strain. The architecture therefore embodies a balance: it seeks to introduce friction where necessary without turning all care into bureaucratic delay. This tension remains ethically important and should not be minimized.

### Privacy and Non-Intrusion

Ethically, privacy is more than data security. In eldercare it is linked to modesty, intimacy, domesticity, and the right not to be excessively monitored. NAI 2.0™ supports privacy by restricting data use, limiting access, and resisting unnecessary expansion of monitoring scope. This aligns with an ethical commitment to non-intrusion as well as legal confidentiality.

### Accountability and Relational Care

Finally, the framework supports accountability by making consequential workflow transitions attributable to identifiable human action. This matters ethically because accountability is part of respectful care. When decisions are made, those affected should not face a situation in which responsibility disappears into system complexity. By preserving explicit human authorization and interruption, NAI 2.0™ helps ensure that care remains answerable to persons rather than diffused across an opaque machine process.

More broadly, the framework supports relational care by refusing the assumption that efficiency should be the dominant value in eldercare AI. It recognizes that older adults often need explanation, patience, contextual attention, and human presence. A governance architecture that protects such space is ethically preferable to one that optimizes only speed and predictive throughput.

Overall, the ethical governance analysis suggests that NAI 2.0™ is not merely compliant in orientation but normatively serious. It embeds values of restraint, recognition, and accountability into architecture. Its strongest ethical contribution is the argument that safe eldercare AI must preserve the conditions of dignified human care, not simply avoid obvious technical failure.

## 6.8 Residual Gaps and Tensions

Despite the alignment strengths identified in this chapter, several residual gaps and tensions remain. A credible thesis must acknowledge these rather than treating governance alignment as complete or frictionless.

First, regulatory mapping is not equivalent to certification. NAI 2.0™ may align conceptually and architecturally with many governance expectations, but actual legal compliance would depend on jurisdiction, implementation details, documentation quality, quality management systems, deployment setting, and regulatory review. The framework is therefore best described as compliance-supportive rather than compliant by default.

Second, implementation quality matters as much as design quality. A well-bounded architecture can still fail in practice if staff are poorly trained, review rights are unclear, infrastructure is weak, or organizational culture discourages pause and challenge. Governance-by-design strengthens control, but it cannot overcome every institutional weakness.

Third, dignity remains difficult to formalize. The framework rightly treats dignity as a system-level requirement, but no architecture can fully capture the relational complexity of dignified care. There is a risk that operationalizing dignity could become reductive if organizations treat scores or checklists as substitutes for genuine attentiveness to the person.

Fourth, there is an ongoing tension between safety-oriented interruption and workflow efficiency. Sacred Pause™ and hardware-enforced review may improve oversight, but they also introduce friction. In under-resourced care environments, such friction could be resisted or resented unless carefully justified and calibrated. A governance architecture that is too cumbersome may invite workaround behavior.

Fifth, legal and policy frameworks continue to evolve. What appears well aligned today may require revision as AI regulation matures, especially in relation to high-risk classification, transparency expectations, and accountability standards. NAI 2.0™ should therefore be understood as adaptable rather than fixed.

These limitations do not invalidate the framework. Rather, they reinforce the central claim of the thesis: that eldercare AI governance must be approached as an ongoing constitutional project involving architecture, regulation, ethics, and institutional practice together. NAI 2.0™ offers a serious foundation for that project, but not its final or universally settled form.

**Concluding Note to Chapter 6**

This chapter has examined NAI 2.0™ as a regulatory, ethical, and governance architecture rather than merely a technical proposal. It has shown that the framework aligns strongly with the logic of governance-by-design by embedding bounded authority, human oversight, interruption, authorization, and traceability into its operational structure. It has also demonstrated meaningful lifecycle alignment across design, development, validation, deployment, monitoring, and change management, indicating that the framework is governable not only at conception but across evolving use.

The analysis further suggests that NAI 2.0™ is highly relevant to a

risk-based SaMD perspective, particularly in relation to intended use clarity, moderate-risk decision support, oversight, documentation, and post-deployment monitoring. Its alignment with the EU AI Act is similarly strong at the level of design principle, especially regarding high-risk sensitivity, human oversight, risk management, and logging. In the domain of data governance, the framework supports the core logic of

HIPAA, GDPR, and CCPA through bounded data use, role-based access, traceability, and minimum-necessary handling. Ethically, the framework is significant because it protects not only safety and accountability but also dignity, autonomy, and relational care.

At the same time, the chapter has acknowledged important limitations. Alignment is not certification, design does not eliminate institutional failure, and dignity cannot be wholly secured through architecture alone. These qualifications are not weaknesses in the argument but signs of academic precision. They clarify that NAI 2.0™ should be understood as a compliance-supportive and ethics-oriented governance model whose full adequacy depends on implementation, validation, and continued review.

With this regulatory and ethical mapping complete, the thesis can now turn to evaluation and feasibility. The next chapter examines what evidence currently supports the framework, how its claims should be interpreted, and what remains to be validated through further empirical and clinical work.

# Chapter 7. Evaluation, Feasibility, and Interpretation of Findings

## 7.1 Introduction

This chapter evaluates the overall strength of NAI 2.0™ as a proposed governance architecture for high-impact eldercare AI deployment. Whereas earlier chapters established the conceptual basis of the framework, described its technical architecture, and mapped its regulatory and ethical alignment, the present chapter addresses a different question: what, at this stage of the thesis, can reasonably be said to have been demonstrated?

This question is methodologically important. A common weakness in AI governance scholarship is the tendency either to overclaim on the basis of conceptual design or, conversely, to dismiss governance architectures because they are not yet supported by large-scale randomized deployment evidence. Both positions are inadequate. In fields such as eldercare AI, where the consequences of premature autonomy may be severe, it is legitimate and necessary to evaluate architectures before full-scale deployment. At the same time, such evaluation must distinguish clearly between what is conceptually coherent, what is architecturally plausible, what is governance-supportive, and what remains empirically unproven.

Accordingly, this chapter does not present NAI 2.0™ as a completed, clinically validated product. Instead, it offers a structured interpretation of the evidence generated across the thesis. The argument advanced is that the framework is strongly supported at the levels of conceptual coherence, governance logic, architectural plausibility, and compliance-oriented design, while its claims regarding real-world efficacy, implementation burden, and measurable outcome improvement remain provisional and dependent on future empirical validation.

The chapter proceeds in eight sections. Section 7.2 restates the evaluation logic used in this thesis and clarifies the standard by which the framework is judged. Section 7.3 evaluates conceptual coherence and internal consistency. Section 7.4 considers architectural and technical feasibility. Section 7.5 examines workflow and operational feasibility in eldercare settings. Section 7.6 reviews the framework's governance and regulatory readiness. Section 7.7 considers dignity, human sovereignty, and the exploratory findings associated with the Elder Dignity Score. Section 7.8 discusses residual risks, unresolved tensions, and interpretive caution. Section 7.9 synthesizes the findings into a balanced overall assessment of the contribution made by the thesis.

The central conclusion of the chapter is that NAI 2.0™ should be regarded as a robust governance prototype at the thesis stage: not yet fully clinically proven, but sufficiently developed in theoretical, architectural, ethical, and regulatory terms to warrant serious consideration as an alternative to more agentic models of AI deployment in eldercare.

## 7.2 Evaluation Logic and Standard of Judgment

The evaluation of NAI 2.0™ is guided by the methodological commitments established in Chapter 4. Because the framework is a governance artefact rather than a single predictive model, it cannot be judged solely by conventional AI metrics such as accuracy, sensitivity, specificity, or latency. Those measures may become relevant in future subsystem validation, but they are insufficient to evaluate a constitutional architecture concerned with bounded authority, human review, interruption, and dignity preservation.

For this reason, the thesis evaluates the framework across five domains:

1. conceptual coherence
2. architectural and technical plausibility
3. workflow and implementation feasibility
4. governance and regulatory alignment
5. ethical and dignity-related adequacy

These domains are not arbitrary. They reflect the actual claims made by the thesis. The framework claims to offer a clearer model of non-agentic eldercare AI than is commonly available. It claims to embed meaningful human sovereignty into architecture rather than rhetoric. It claims to support safety and dignity by introducing boundedness, pause, override, and traceability into workflow. And it claims to be more legible to emerging regulatory expectations than architectures that treat autonomy expansion as the default trajectory of innovation.

The appropriate standard of judgment is therefore not whether the framework has already transformed eldercare outcomes across multiple clinical sites. That would require a far more mature stage of development. Rather, the relevant question is whether the thesis has provided sufficient evidence to justify the following more modest but important conclusions:

— that the framework is internally coherent;

— that its core components can be meaningfully related to real governance problems;

— that its workflow logic is operationally intelligible;

— that it addresses regulatory and ethical concerns in a structured way;

— and that it identifies its own limitations without collapsing into self-contradiction.

In this chapter, the phrase supported is therefore used carefully. It means that the available conceptual, design, governance, and feasibility evidence provides a reasonable basis for advancing the claim in question. It does not mean that the claim has been conclusively verified in large-scale practice. This distinction is central to the intellectual honesty of the thesis.

## 7.3 Conceptual Coherence and Internal Consistency

The first and strongest area of support for NAI 2.0™ lies in its conceptual coherence. One of the thesis's key contributions is that it does not merely use terms such as non-agentic AI, human sovereignty, constitutional architecture, and dignity preservation as rhetorical slogans. Instead, it defines them in a structured and interdependent way.

The framework is conceptually coherent in at least four important respects.

### 7.3.1 Clarity of the Core Problem

First, the thesis identifies a clear and credible problem: in eldercare, AI systems may become practically authoritative even where humans are formally retained in the loop. This problem is plausible and analytically significant. The concern is not simply that AI might make errors, but that system design may permit recommendations to acquire decision-like force through workflow speed, interface bias, institutional pressure, or inadequate interruptibility. This problem framing is one of the intellectual strengths of the thesis because it moves beyond narrow debates about model performance and focuses on the structure of authority itself.

### 7.3.2 Consistency of the Non-Agentic Position

Second, the framework is internally consistent in its commitment to non-agentic design. The thesis does not merely announce a preference for human oversight while quietly relying on machine-led progression. Instead, the entire architecture is built around the claim that high-impact outputs should remain review-dependent, bounded, and interruptible. The data layer, inference layer, constitutional constraints, hardware-enforced controls, review mechanisms, and logging functions all reflect this same foundational logic.

This coherence matters because many AI governance proposals fail precisely at this point. They endorse human-centered values in principle while leaving the operational architecture strongly tilted toward autonomy expansion. NAI 2.0™ avoids much of this inconsistency by ensuring that the language of boundedness is matched by actual architectural restraint.

### 7.3.3 Integration of Dignity with Governance

Third, the framework is conceptually distinctive in how it integrates

dignity into system governance. Dignity is not treated as a decorative ethical add-on. It is linked directly to pause, explanation, privacy proportionality, refusal space, and preservation of human interpretive presence. This integration is significant because eldercare AI is often discussed through safety and efficiency alone, leaving dignity underdeveloped despite its obvious importance in care settings involving dependency and vulnerability.

The thesis is careful not to claim that dignity can be fully formalized. That caution strengthens rather than weakens the argument. The conceptual contribution lies in showing that dignity can still function as a legitimate design requirement even if it resists total quantification.

### 7.3.4 Logical Relationship Between Components

Fourth, the four constitutional elements---3ZEROS Sanctuary,

Sacred Pause™, Sovereign Brake, and Tiger .1x Key™---are logically related rather than merely branded. Each performs a distinct governance role:

— 3ZEROS Sanctuary establishes the protected operating

environment;

— Sacred Pause™ creates non-bypassable review time for high-impact

outputs;

— Sovereign Brake preserves the power to halt progression;

— Tiger .1x Key™ ensures explicit human authorization for bounded

transition.

The interaction among these components is conceptually intelligible. Together they form a layered response to the central problem of authority drift. The framework is therefore

stronger than proposals that rely on a single vague commitment to "human oversight" without specifying how that oversight becomes real.

Taken together, these observations support the conclusion that NAI 2.0™ is highly robust at the conceptual level. Its central claims are mutually reinforcing rather than contradictory, and its major terms have been defined with sufficient precision to sustain scholarly analysis.

## 7.4 Architectural and Technical Feasibility

The second domain of evaluation concerns whether the framework is architecturally plausible. At this stage of the thesis, plausibility is more important than proof of full implementation success. The question is whether the proposed design could reasonably be instantiated in a real eldercare environment without collapsing into technical incoherence.

The evidence presented in Chapters 5 and 6 suggests that the answer is yes, with important qualifications.

### 7.4.1 Plausibility of the Layered Architecture

The seven-layer architecture described in Chapter 5 is technically credible. Data intake, bounded inference, constitutional rule checking, human review interfaces, logging, and fallback procedures are all familiar architectural functions in complex digital systems. What is distinctive here is not the existence of these functions individually, but their organization around non-agentic governance.

There is nothing intrinsically implausible about such a layered structure. On the contrary, the architecture benefits from modularity. Because each layer performs a defined role, the design allows for clearer reasoning about responsibility, error pathways, and change management. This is particularly useful in governance-sensitive settings where system opacity can undermine accountability.

### 7.4.2 Feasibility of Bounded Outputs

The framework's insistence that outputs remain bounded and non-self-executing is also technically feasible. Many existing systems already distinguish between informative prompts and executive actions. NAI 2.0™ extends this distinction by making it constitutionally central. There is no technical necessity requiring eldercare AI to be fully agentic. The decision to bind outputs to review is therefore a governance choice rather than a capability limitation.

This finding is important because it undermines a common assumption that increasing sophistication must naturally lead to increasing autonomy. The thesis demonstrates that a powerful system can remain intentionally constrained.

### 7.4.3 Hardware-Enforced Control Logic

The most distinctive technical feature of the framework is its use of

hardware-enforced safety logic. Here the evidence is more mixed, though still supportive. The concept itself is plausible: healthcare environments already use physical or device-linked controls in other safety-critical contexts. Requiring explicit materially effective authorization for certain transitions is not technically unrealistic.

However, the exact implementation of hardware enforcement will vary significantly across institutions, infrastructure maturity, and cost conditions. This means that while the principle is credible, its operational form remains partly implementation-dependent. The thesis is strongest when it argues for the need for stronger-than-software controls in high-impact contexts, and somewhat less strong when implying any one universal hardware solution.

### 7.4.4 Technical Strengths and Remaining Unknowns

The major technical strengths of the framework are therefore:

— architectural modularity;

— clear separation between inference and action;

— strong interruptibility logic;

— traceability of consequential transitions;

— compatibility with role-based access and privacy-conscious design.

The major unknowns are:

— the precise burden introduced by hardware-linked controls;

— interoperability with existing eldercare information systems;

— calibration of thresholds for pause and escalation;

— the degree to which different institutions could implement the same

constitutional logic consistently.

These unknowns do not invalidate the architecture. They simply define the boundary between design plausibility and deployment proof. On balance, the thesis supports the conclusion that NAI 2.0™ is

architecturally feasible in principle, though not yet empirically optimized across real institutional environments.

## 7.5 Workflow and Operational Feasibility in Eldercare

A framework may be conceptually elegant and technically possible yet still fail in practice if it does not fit care workflows. This is especially true in eldercare, where staffing levels, cognitive load, relational demands, documentation pressures, and emotional complexity all shape what can realistically be sustained.

The thesis evaluates workflow feasibility through scenario logic, role mapping, escalation analysis, and the broader sociotechnical reasoning set out in earlier chapters. This evaluation suggests a nuanced conclusion: NAI 2.0™ appears operationally feasible, but only if deployed with careful calibration and cultural support.

### 7.5.1 Strength: The Architecture Matches High-Impact Care Logic

One of the framework's strongest operational features is that it does not impose maximum friction everywhere. Instead, it distinguishes between low-impact, moderate-impact, and high-impact outputs. This is essential. If every AI output required the same degree of formal review, the system would likely become unusable. By reserving Sacred Pause™ and stronger controls for outputs with significant safety, privacy, autonomy, or dignity implications, the framework shows sensitivity to actual workflow conditions.

This stratified governance logic makes practical sense in eldercare. Many care-related tasks are routine and can tolerate lighter-touch support. Others are consequential and require deliberate pause. The framework's ability to differentiate these cases is a major feasibility strength.

7.5.2 Strength: Human Review Is Positioned Before Consequential Action

The review architecture is also operationally meaningful because it places human intervention before consequential progression rather than after. This improves practical oversight. Retrospective review is useful for audit, but it cannot prevent harm that has already occurred. By locating pause and authorization at the point of transition, the system aligns workflow with the actual timing of ethical significance.

This feature strengthens the thesis's claim that human sovereignty is operational rather than symbolic.

### 7.5.3 Tension: Friction Versus Efficiency

The main operational challenge is the obvious one: friction. Sacred Pause™, bounded authorization, and hardware-linked control are valuable because they slow consequential progression. Yet the very friction that protects against over-automation may be experienced by staff as burden, delay, or interruption. In under-resourced care environments, this could create resistance.

The thesis does not ignore this problem, and that is one of its strengths. Rather than pretending that all friction is good, it argues for disciplined friction---that is, friction introduced only where the risk of unreflective progression justifies it. This is a defensible position, but it remains a point requiring empirical calibration. Institutions differ, and what feels proportionate in one setting may feel excessive in another.

### 7.5.4 Tension: Training and Culture Dependence

Operational feasibility also depends heavily on institutional culture. A system built around meaningful interruption requires a care environment in which pausing is seen as legitimate rather than inefficient. If staff are pressured to prioritize speed above all else, even well-designed safeguards may be informally weakened or resented.

This means that NAI 2.0™ is best understood as a sociotechnical intervention, not merely a technical deployment. Its feasibility depends on training, leadership, governance clarity, and cultural reinforcement. This requirement is demanding, but not unusual in safety-critical care systems.

### 7.5.5 Overall Operational Assessment

Overall, the workflow analysis supports a cautious but positive conclusion. The framework appears feasible because it is structured, role-sensitive, and aware of operational burden. However, its success in practice will depend on thoughtful threshold setting, realistic staffing assumptions, and institutional willingness to treat pause and override as legitimate features of safe care. In short, the framework is operationally plausible, but not operationally self-sufficient.

## 7.6 Governance and Regulatory Readiness

The regulatory and governance mapping conducted in Chapter 6 provides one of the most persuasive dimensions of support for NAI 2.0™. The framework does not merely declare itself ethical or compliant; it demonstrates structured alignment with contemporary expectations relating to risk-sensitive AI deployment in healthcare.

### 7.6.1 Strength: Clarity of Intended Use

A major governance strength of the framework is its clear intended use. NAI 2.0™ is not positioned as a general autonomous care intelligence. It is a bounded eldercare governance architecture for high-impact decision support. This clarity matters because regulatory systems depend heavily on scope definition. A narrowly bounded system is easier to classify, document, monitor, and govern than one with diffuse or expanding authority.

### 7.6.2 Strength: Human Oversight Is Architecturally Embedded

Another major strength is the framework's treatment of human oversight. Current AI governance increasingly distinguishes between nominal oversight and meaningful oversight. NAI 2.0™ aligns strongly with the latter concept because it gives humans not only formal responsibility but also structural power: the ability to pause, stop, authorize, and review before consequential progression occurs.

This is one of the thesis's most important achievements. It makes the framework legible to emerging regulatory expectations around high-risk AI, medical decision support, and accountable system operation.

### 7.6.3 Strength: Traceability and Auditability

The architecture's logging and authorization features also support governance readiness. Inputs, classifications, pause events, authorizations, and overrides can be recorded and reviewed. This enables incident analysis, quality improvement, and post-deployment governance. In compliance-sensitive settings, such traceability is essential.

### 7.6.4 Strength: Data Governance Discipline

The framework also demonstrates maturity in its approach to privacy and data handling. By emphasizing minimum necessary intake, role-based access, bounded purpose, and controlled traceability, NAI 2.0™ aligns with the general logic of modern data governance. This is particularly important in eldercare, where privacy harm may also be dignity harm.

### 7.6.5 Limitation: Readiness Is Not Certification

The main limitation here is one the thesis openly acknowledges:

regulatory readiness is not regulatory approval. Formal compliance would require jurisdiction-specific implementation, quality management processes, documentation, legal review, and perhaps conformity assessment depending on use case. The thesis does not claim to have completed these processes.

This limitation is not fatal. On the contrary, it strengthens the credibility of the analysis by distinguishing between alignment and

certification. The appropriate conclusion is therefore that NAI 2.0™ is highly governance-ready in design, even though it is not yet proven as a fully certified deployment product.

## 7.7 Dignity, Human Sovereignty, and Exploratory Assessment Findings

One of the distinctive claims of the thesis is that eldercare AI must be judged not only by safety and compliance but also by its implications for dignity and human sovereignty. This is where NAI 2.0™ attempts to make its most original normative contribution.

### 7.7.1 Dignity as an Evaluative Domain

The framework performs well conceptually in this domain because it identifies concrete architectural conditions that are relevant to dignified care: explanation, pause, non-coercion, privacy proportionality, and continued human presence in consequential review. This is stronger than simply asserting that dignity matters. It shows how dignity can influence the design of workflow.

### 7.7.2 Human Sovereignty as a Structural Reality

The framework also supports human sovereignty more robustly than many conventional decision-support systems. Sovereignty, in the thesis, does not mean vague human superiority in principle. It means that humans retain materially effective power over whether the system's recommendations alter care trajectories. Sacred Pause™, Sovereign Brake, and Tiger .1x Key™ make this claim structurally credible.

This is an important evaluative finding. If the thesis had merely argued that humans should remain responsible, it would add little to existing governance language. Its stronger contribution is showing how sovereignty can be operationalized.

### 7.7.3 Exploratory Use of the Elder Dignity Score

The Elder Dignity Score, as introduced in Chapter 4, is used here cautiously as an exploratory lens rather than as a validated psychometric conclusion. Within the logic of the thesis, the score indicates that the framework performs positively on dimensions such as:

— opportunity for pause before consequential action;

— preservation of human interpretive involvement;

— support for explanation rather than silent progression;

— resistance to coercive machine-led workflow;

— privacy-conscious boundedness;

— visible accountability.

These are not trivial findings. They suggest that the framework is better positioned than more seamless automation models to preserve conditions associated with dignified care.

However, the limitations of the instrument are equally important. Dignity is context-sensitive and relational. No score can exhaust its meaning. Moreover, user experience in real care settings may differ from design expectations. The Elder Dignity Score therefore strengthens the thesis only when interpreted modestly: as evidence that dignity has been systematically considered in evaluation, not as proof that dignity is guaranteed.

### 7.7.4 Overall Ethical Interpretation

Overall, the findings in this domain are favorable. NAI 2.0™ appears to preserve dignity and human sovereignty more deliberately than architectures that prioritize continuous automation. Its ethical value lies in restraint. It creates formal room for hesitation, challenge, and contextual recognition. In eldercare, where depersonalization is a persistent risk, this is a major strength.

## 7.8 Residual Risks, Unresolved Tensions, and Interpretive Caution

No serious evaluation would be complete without considering what the framework does not yet prove or where it may encounter difficulties. Several unresolved issues remain.

### 7.8.1 Risk of Over-Burdening Care Workflows

The first risk is that protective friction could become excessive. If pause conditions are triggered too often, if authorization pathways are too cumbersome, or if hardware-linked controls are poorly integrated, users may experience the system as obstructive. This could generate workarounds, resentment, or superficial compliance. The thesis addresses this risk conceptually through stratified governance, but only future deployment studies can show where the right balance lies.

### 7.8.2 Risk of Symbolic Rather Than Lived Dignity

A second risk is that the language of dignity could become formalized without fully shaping practice. It is possible for organizations to adopt dignity-oriented language while still operating in rushed or impersonal ways. The thesis reduces this risk by linking dignity to specific architectural conditions, but it cannot fully eliminate the possibility of symbolic adoption.

### 7.8.3 Risk of Organizational Misfit

A third issue is institutional variation. Eldercare settings differ widely in staffing, digital maturity, funding, legal environment, and care philosophy. A framework that is feasible in one environment may require substantial adaptation in another. NAI 2.0™ is therefore better understood as a governance model with transfer potential rather than a universal plug-and-play template.

### 7.8.4 Risk of Partial Validation Being Overread

A fourth concern is interpretive overreach. Because the framework performs well conceptually and regulatorily, there is a temptation to treat it as already validated in practical terms. That would be mistaken. The thesis supports claims of coherence, plausibility, and readiness. It does not yet support strong claims of superior outcome performance, cost-effectiveness, or universal user acceptance.

### 7.8.5 The Need for Future Empirical Work

These limitations point directly to the next stage of research. Future work should include:

— controlled pilot implementations in eldercare settings;

— workflow burden studies;

— role-specific usability testing;

— validation of pause thresholds and override patterns;

— refinement and validation of dignity-related assessment tools;

— comparative analysis against more conventional AI deployment models.

These requirements do not diminish the present thesis. They define the proper path from governance prototype to mature deployment evidence.

## 7.9 Overall Interpretation of the Findings

When the evidence from the entire thesis is considered together, a balanced interpretation becomes possible.

First, NAI 2.0™ succeeds strongly as a conceptual and governance contribution. It offers a more precise vocabulary for thinking about bounded AI in eldercare and advances the important claim that the central design question is not only what AI can do, but what authority it should be allowed to exercise.

Second, the framework succeeds substantially as an architectural proposal. Its layered design, bounded outputs, interruptibility logic, traceability, and role-sensitive authorization are all technically plausible and mutually reinforcing. It demonstrates that non-agentic AI can be specified in operational rather than merely philosophical terms.

Third, the framework is persuasive as a compliance-supportive and regulation-aware model. Its alignment with risk-sensitive healthcare governance, human oversight expectations, privacy principles, and lifecycle accountability is a major strength of the thesis. This gives the work practical relevance beyond abstract ethics.

Fourth, the framework makes a meaningful ethical contribution by centering dignity and sovereignty. In doing so, it addresses dimensions of eldercare often overlooked in AI system design. Its insistence on pause, explanation, and non-coercive progression is normatively significant.

Fifth, the framework remains pre-empirical in its strongest claims about real-world performance. This is not a flaw so long as it is acknowledged clearly. The thesis should be read as establishing the legitimacy, coherence, and feasibility of the framework, while also identifying the empirical agenda required for further validation.

On that basis, the most defensible overall judgment is that NAI 2.0™ constitutes a substantial doctoral-level contribution in the form of a carefully reasoned governance architecture for eldercare AI. It does not claim to end the debate or complete the empirical task. What it does achieve is the articulation of a serious alternative to autonomy-driven AI deployment: one that is architecturally bounded, ethically serious, regulatorily legible, and specifically responsive to the vulnerabilities of eldercare.

**Concluding Note to Chapter 7**

This chapter has evaluated NAI 2.0™ across the principal domains relevant to its claims: conceptual coherence, technical plausibility, workflow feasibility, governance readiness, and dignity-oriented ethical adequacy. The analysis has shown that the framework is strongest where the thesis intended it to be strongest: as a constitutional design model that preserves meaningful human sovereignty over high-impact eldercare AI.

The findings suggest that NAI 2.0™ is not merely a rhetorical call for responsible AI. It is a structured and credible architecture that translates responsibility into bounded outputs, enforced pause, explicit authorization, interruptibility, and traceable accountability. These features make it especially relevant to eldercare, where the risks of depersonalization, over-automation, and dignity erosion are unusually significant.

At the same time, the chapter has maintained appropriate caution. The framework is not yet validated through large-scale clinical deployment, and its operational burden, institutional fit, and measurable outcome advantages remain matters for future research. This does not weaken the contribution. Rather, it clarifies it. The thesis has provided a strong foundation on which empirical validation can proceed without confusion about the framework's purpose or claims.

With the evaluation now complete, the thesis is positioned to move into its final integrative stage. The next chapter draws together the overall argument, restates the core contribution of the study, identifies its implications for eldercare AI governance, and sets out a future research and implementation agenda.

# Chapter 8. Conclusion and Future Directions

## 8.1 Introduction

This thesis set out to address a pressing and insufficiently resolved problem in contemporary eldercare innovation: how can artificial intelligence be used in high-impact care settings without allowing machine outputs to acquire unbounded practical authority over vulnerable persons and dignity-sensitive care processes? The study argued that this is not merely a technical question of predictive performance, nor solely an ethical question of abstract principle. It is a problem of

governance architecture. Specifically, it is a problem of how authority is distributed, how consequential transitions are controlled, and how meaningful human sovereignty can be preserved when AI systems enter real care workflows.

In response, the thesis developed NAI 2.0™ as a hardware-enforced, non-agentic AI governance framework for eldercare. Rather than assuming that AI capability should naturally progress toward greater autonomy, the framework was designed around the opposite premise: in eldercare, value may lie in disciplined limitation. Where older adults may be exposed to dependency, frailty, cognitive fluctuation, surveillance risk, or diminished capacity to challenge institutional systems, the burden falls on designers and deployers to ensure that AI remains bounded, review-dependent, interruptible, and traceable.

The core contribution of the study has therefore been the articulation and evaluation of a constitutional approach to eldercare AI. This approach embeds governance not only in policy statements or professional guidance, but in the operational architecture of the system itself. Through the interlocking mechanisms of 3ZEROS Sanctuary, Sacred Pause™, Sovereign Brake, and Tiger .1x Key™, the framework seeks to ensure that consequential AI outputs do not move seamlessly into action without explicit human review and authorization.

The thesis does not claim that NAI 2.0™ is a fully validated clinical product, nor that it resolves all challenges of AI deployment in care. Its claims have been deliberately more disciplined. It has argued that the framework is conceptually coherent, architecturally plausible, ethically serious, workflow-aware, and

substantially aligned with emerging governance expectations. It has also acknowledged that questions of real-world efficacy, implementation burden, user acceptance, and institutional fit remain matters for future empirical investigation.

This final chapter draws the thesis together. It restates the central problem and rationale of the study, synthesizes the principal findings across the preceding chapters, identifies the thesis's original contributions, discusses its practical and policy implications, acknowledges its limitations, and sets out a future research and implementation agenda. The chapter concludes by reaffirming the broader claim that eldercare AI should be

approached not as a domain for maximal automation, but as a domain requiring constitutional restraint, human sovereignty, and dignity-preserving design.

## 8.2 Restatement of the Research Problem and Purpose

The starting point of the thesis was the observation that contemporary AI discourse often treats human oversight as a sufficient answer to the risks of high-impact automation. In healthcare and care settings, systems are commonly described as decision support tools, advisory systems, or assistive technologies, thereby implying that ultimate control remains securely with human professionals. However, the literature reviewed in Chapter 2 and the conceptual analysis undertaken in Chapter 3 demonstrated that this assumption is frequently unstable. Systems need not be formally autonomous in order to become practically authoritative. Through workflow integration, interface design, standardization pressure, institutional trust, or time-constrained environments, machine outputs may come to guide decisions in ways that are difficult to contest or slow down.

This problem is particularly acute in eldercare. Older adults often occupy positions of heightened situational vulnerability. Care decisions in this domain may affect safety, bodily wellbeing, privacy, autonomy, daily routine, emotional security, and dignity. Eldercare is also distinguished by the relational and longitudinal nature of care itself. Unlike some acute healthcare contexts, eldercare often unfolds through ongoing dependency relationships, repeated interpretation, negotiated support, and close contact with intimate dimensions of life. As a result, the consequences of poorly bounded AI are not limited to technical error. They also include depersonalization, coercive workflow, surveillance creep, diminished refusal space, and the erosion of human attentiveness.

The thesis therefore pursued a specific purpose: to develop and evaluate a governance architecture that could preserve meaningful human sovereignty over high-impact eldercare AI processes while remaining technically plausible and regulatorily legible. This required moving beyond broad ethical aspiration toward design-level specification. The study asked not simply whether AI should be used responsibly, but how responsibility could be architecturally realized. It also asked how dignity and sovereignty could be treated not as rhetorical values but as operational constraints.

To achieve that purpose, the thesis took an interdisciplinary approach, drawing together healthcare AI literature, eldercare ethics, safety engineering, sociotechnical systems thinking, and regulatory governance analysis. The result was the formulation of NAI 2.0™ as a non-agentic constitutional framework intended to keep AI supportive without allowing it to become sovereign.

## 8.3 Synthesis of the Major Findings

The findings of the thesis may be synthesized across five main domains: literature and problem identification, conceptual development, architectural design, governance and regulatory alignment, and evaluative interpretation.

### 8.3.1 Literature and Problem Identification

The literature review established that while there is substantial scholarship on AI in healthcare, digital technologies in eldercare, human oversight, privacy, dignity, and AI regulation, these bodies of work remain fragmented. Existing literature provides valuable insights into model performance, workflow integration, automation bias, high-risk governance, and ethical concerns, yet it does not offer a sufficiently integrated eldercare-specific governance architecture for bounded AI authority.

Several gaps were identified. First, human oversight is widely invoked but often under-specified. Second, eldercare is frequently treated as a subset of generic healthcare or digital care innovation, rather than as a distinct governance context marked by heightened dignity sensitivity and relational care concerns. Third, dignity is often acknowledged at the level of principle without being translated into architectural conditions. Fourth, current governance literature speaks powerfully about accountability, transparency, and traceability, but less clearly about how high-impact outputs should be made non-bypassably interruptible within workflow. Finally, hardware-enforced governance mechanisms are largely absent from contemporary eldercare AI discussion despite their relevance in other safety-critical domains.

The literature review therefore justified the thesis's central intervention: the need for a framework that integrates bounded authority, meaningful human control, privacy-conscious design, and dignity preservation into a single architectural logic.

### 8.3.2 Conceptual Development

The thesis then developed a conceptual vocabulary capable of supporting that intervention. Chapter 3 clarified the meanings of non-agentic AI, human sovereignty, constitutional architecture, and

dignity preservation. These concepts were not treated as aspirational labels, but as analytically distinct and interdependent elements of a governance model.

The concept of non-agentic AI was particularly important. Rather than defining AI merely by capability, the thesis defined non-agentic AI as AI whose outputs remain bounded, review-dependent, and non-self-executing in relation to consequential care processes. This conceptual move shifted attention from intelligence as performance to intelligence as governed participation within a care system.

Similarly, human sovereignty was defined not as symbolic final responsibility but as materially effective authority: the real power to pause, reject, review, authorize, or redirect

AI-influenced pathways before consequential action occurs. This distinction was essential because it exposed the inadequacy of nominal human-in-the-loop models where humans remain present but insufficiently empowered.

The thesis also advanced the idea of constitutional architecture, meaning a design in which system constraints are embedded structurally rather than delegated solely to policy, training, or user intention. This concept linked the normative concerns of the study to its design ambitions. It made clear that if dignity, accountability, and bounded authority are to be reliable, they must shape the permitted operational logic of the system itself.

Finally, the thesis conceptualized dignity preservation as a design-relevant requirement. It argued that dignity in eldercare is linked to explanation, pause, refusal space, contextual review, privacy proportionality, and the continued presence of human interpretive judgment. In doing so, it offered a bridge between ethical reflection and system design.

### 8.3.3 Architectural Design and System Logic

Chapter 5 translated these concepts into the proposed NAI 2.0™ architecture. The framework was presented as a layered and bounded system for high-impact eldercare AI deployment. Its architecture intentionally separated inference from action and prevented the system from autonomously progressing into consequential outcomes.

The four core constitutional components of the framework were central to this architecture:

— 3ZEROS Sanctuary established the protected and bounded operational environment within which the system functions.

— Sacred Pause™ created a non-bypassable period of review for outputs classified as high impact.

— Sovereign Brake preserved the ability of authorized humans to halt or interrupt progression when uncertainty, concern, or contextual mismatch arises.

— Tiger .1x Key™ operationalized explicit human authorization for transitions that should not occur automatically.

Together, these mechanisms formed a governance model designed to resist authority drift. Instead of assuming that oversight would emerge through good practice alone, the framework created structural conditions in which certain outputs could not become consequential without human intervention.

The system design also included privacy-sensitive data handling, traceability, role-based authorization logic, and hardware-reinforced safety boundaries. These features demonstrated that bounded AI need not be conceptually vague. It can be architected in a manner that is operationally intelligible and technically plausible.

### 8.3.4 Governance and Regulatory Alignment

Chapter 6 examined the framework through the lens of regulatory and governance expectations. One of the most significant findings of the thesis was that NAI 2.0™ is strongly aligned, at the design level, with the broader logic of emerging high-risk AI governance. This includes expectations around human oversight, traceability, risk-sensitive deployment, intended use clarity, auditability, privacy-aware operation, and lifecycle accountability.

The framework was found to be particularly strong in three respects. First, it gives human oversight real structural force. Second, it enables audit and review through traceable authorization and logging pathways. Third, it integrates privacy and bounded-purpose logic into the architecture rather than treating them as external compliance issues.

At the same time, the thesis was careful to distinguish governance readiness from formal certification. The study did not claim regulatory approval or legal clearance for any specific jurisdictional deployment. Rather, it argued that the framework is designed in a manner that is more legible to governance and regulatory assessment than architectures that treat autonomy expansion as the default model.

### 8.3.5 Evaluation and Interpretation

Chapter 7 provided the overall evaluative interpretation of the thesis. The main conclusion was that NAI 2.0™ is best understood as a robust governance prototype. The framework is strongly supported in terms of conceptual coherence, design logic, governance relevance, and ethical seriousness. It offers a credible and carefully reasoned alternative to autonomy-oriented eldercare AI models.

However, the thesis also recognized that its strongest claims are pre-empirical. It does not yet prove that the framework improves patient outcomes, reduces incidents, lowers cost, or maximizes user satisfaction across real care environments. Those questions require future pilot deployment, usability testing, institutional comparison, and implementation research.

The thesis therefore concluded with a deliberately balanced judgment. NAI 2.0™ is not a finished clinical product, but it is far more than a rhetorical proposal. It is a mature and substantial doctoral-level contribution in the form of a constitutional AI governance architecture for eldercare.

## 8.4 Direct Answers to the Research Questions

This section draws together the findings in direct relation to the questions that guided the study.

### 8.4.1 Main Research Question

How can a hardware-enforced, non-agentic AI governance architecture preserve meaningful human sovereignty and dignity in high-impact eldercare settings while remaining technically plausible and regulatorily aligned?

The thesis has shown that this can be done by designing AI systems in which consequential outputs are bounded, interruptible, review-dependent, and explicitly authorized before progression. Human sovereignty is preserved not through symbolic responsibility alone, but through materially effective control points built into the architecture. Dignity is supported by maintaining pause, contextual review, explanation, privacy proportionality, and resistance to seamless machine-led progression. Technical plausibility is achieved through layered modular architecture, separation of inference from action, and hardware-linked safety logic. Regulatory alignment is supported through intended-use clarity, traceability, auditability, role-based governance, and privacy-sensitive design.

### 8.4.2 Sub-Question 1

What shortcomings exist in current AI, healthcare, and eldercare governance literature regarding bounded authority and meaningful human control?

The literature reveals significant shortcomings in the specification of human oversight, the translation of ethical principles into operational design, the treatment of eldercare as a distinct governance context, and the management of authority drift. Current discourse often assumes that human presence equals human control, and that policy-level commitments are sufficient to govern machine influence. This thesis demonstrated that these assumptions are inadequate in high-impact eldercare environments.

### 8.4.3 Sub-Question 2

What conceptual foundations are necessary to define a non-agentic AI architecture suitable for eldercare?

The study showed that four foundations are central: non-agentic AI, human sovereignty, constitutional architecture, and dignity preservation. These concepts make it possible to define a governance model in which AI remains supportively intelligent without becoming

practically sovereign. They also allow eldercare-specific ethical concerns to be integrated into design requirements.

### 8.4.4 Sub-Question 3

How can governance principles such as pause, interruption, review, and explicit authorization be embedded into technical architecture rather than left at the level of policy alone?

The thesis answered this by developing the NAI 2.0™ constitutional components and layered system model. Pause is embedded through Sacred Pause™. Interruption is embedded through Sovereign Brake. Authorization is embedded through Tiger .1x Key™. Protected operating boundaries are embedded through 3ZEROS Sanctuary. Together these mechanisms convert governance principles into architecture.

### 8.4.5 Sub-Question 4

To what extent does the proposed NAI 2.0™ framework align with regulatory, privacy, and ethical expectations relevant to healthcare and high-risk AI deployment?

The evaluation found substantial alignment at the design level. The framework supports the logic of risk-sensitive oversight, meaningful human control, data minimization, auditability, role-based governance, and lifecycle accountability. It is therefore governance-ready in conceptual and architectural terms, though not yet formally certified or deployed at scale.

### 8.4.6 Sub-Question 5

How feasible is the framework as a workflow-sensitive model for real eldercare settings, and what tensions or limitations remain unresolved?

The framework appears operationally plausible, especially because it uses stratified governance rather than applying maximal friction to all outputs. However, practical tensions remain around workflow burden, cultural acceptance of pause, interoperability, training demands, and the risk that safeguards may be experienced as obstructive if poorly calibrated. These issues define the agenda for future empirical testing rather than negating the framework's value.

## 8.5 Original Contributions of the Thesis

The thesis makes several original contributions to knowledge and practice.

### 8.5.1 A New Eldercare-Specific Governance Architecture

The most obvious contribution is the development of NAI 2.0™ itself as a distinct eldercare-specific AI governance architecture. While existing literature discusses AI ethics and oversight in general terms, this thesis offers a concrete model designed specifically for high-impact eldercare use.

8.5.2 Conceptual Clarification of Non-Agentic AI and Human Sovereignty

A second major contribution is conceptual. The thesis clarifies the meaning of non-agentic AI and human sovereignty in a way that is operationally useful. This helps move discourse beyond vague references to responsible or human-centered AI.

### 8.5.3 A Constitutional Approach to AI Design

Third, the study contributes the notion of constitutional AI architecture in eldercare. This is significant because it reorients AI governance away from permissive innovation constrained only after the fact and toward structurally bounded design from the outset.

### 8.5.4 Translation of Dignity Into Design Logic

Fourth, the thesis contributes to eldercare ethics by showing how dignity can be linked to architectural features such as pause, explanation, privacy proportionality, and preserved human review. This is an important step beyond treating dignity as ethically important but technically untranslatable.

8.5.5 Introduction of Hardware-Enforced Governance Into Eldercare AI Debate

Fifth, the study extends existing discussion by introducing

hardware-enforced governance logic into eldercare AI scholarship. This is particularly novel because it draws from high-reliability safety thinking while adapting it to a dignity-sensitive care context.

### 8.5.6 A Governance-Oriented Evaluation Framework

Finally, the thesis contributes methodologically by evaluating the framework across multiple domains: conceptual coherence, architectural plausibility, workflow feasibility, governance alignment, and dignity-related adequacy. This multi-domain evaluative logic is itself useful for future governance-oriented AI scholarship.

## 8.6 Implications for Theory, Practice, and Policy

### 8.6.1 Implications for Theory

Theoretically, the study suggests that AI governance in care should be analyzed as a problem of authority design rather than simply capability control. This is an important shift. It means that future scholarship should attend more carefully to how systems become influential in practice, not merely to whether they are formally autonomous. It also suggests that eldercare deserves more specific treatment as a governance context with its own conceptual demands.

The thesis further implies that dignity can and should play a more central role in AI architecture discussions. Rather than being treated as too vague for design, dignity may function as a constraint that shapes system timing, escalation pathways, privacy boundaries, and review structures.

### 8.6.2 Implications for Practice

For designers and implementers, the thesis suggests that high-impact eldercare AI should not be built around seamless automation. Instead, systems should differentiate between low-impact support and high-impact recommendation, reserving stronger review and authorization requirements for the latter. This is a practical design principle with broad relevance.

The thesis also implies that meaningful human control must be made easy to exercise. Staff should be able not only to see outputs but to pause, question, reject, and redirect them without unreasonable burden. In practical terms, this means interface design, alert sequencing, role permissions, and escalation protocols all matter as much as model accuracy.

### 8.6.3 Implications for Policy and Regulation

For policymakers and regulators, the study suggests that governance frameworks should look beyond declarations of human oversight and examine whether control is architecturally real. A system that nominally includes human review but operationally privileges machine-led progression may not satisfy the spirit of meaningful oversight even if it satisfies minimal formal requirements.

The thesis also suggests that eldercare may require more specific governance attention within the broader field of health-related AI regulation. Regulatory models that distinguish

by risk level should also attend to the distinctive vulnerabilities and relational harms associated with long-term care and support environments.

## 8.7 Limitations of the Study

No doctoral thesis can resolve every dimension of a complex sociotechnical problem, and this study has several limitations that should be acknowledged clearly.

First, the thesis is primarily a design and evaluation study, not a large-scale empirical implementation study. While it provides strong conceptual and governance analysis, it does not yet offer real-world deployment data across multiple eldercare institutions.

Second, the technical architecture is presented as a plausible and structured model, but not as a fully engineered commercial or clinical product. Some implementation details, especially regarding hardware enforcement and interoperability, will necessarily vary by institutional context.

Third, although the thesis has engaged seriously with privacy, dignity, and human control, it has not exhaustively addressed every issue relevant to AI governance, such as cross-cultural variation in eldercare norms, reimbursement structures, procurement politics, or all possible dimensions of algorithmic bias. These remain important and deserve further work.

Fourth, the Elder Dignity Score was used cautiously as an exploratory evaluation lens rather than as a validated psychometric measure. It helped structure evaluation, but it should not be interpreted as definitive proof of preserved dignity in real care settings.

Fifth, because the framework is intentionally bounded and friction-sensitive, there is an unresolved tension between protection and burden. The thesis argues that disciplined friction is justified for high-impact outputs, but only empirical use can establish the most proportionate threshold settings in practice.

These limitations do not undermine the thesis's contribution. Rather, they define its proper status as a serious governance prototype and as a foundation for future validation.

## 8.8 Future Research Agenda

A central outcome of the thesis is that it opens a substantial and important research agenda. Several directions are especially urgent.

### 8.8.1 Pilot Implementation Studies

The most immediate need is for controlled pilot implementations of the NAI 2.0™ framework in eldercare settings. Such studies should examine how the architecture functions in practice, how users respond to pause and authorization mechanisms, and whether the framework can be integrated without unacceptable workflow burden.

### 8.8.2 Usability and Human Factors Testing

Future research should include rigorous human factors and usability studies involving nurses, care staff, clinicians, administrators, older adults, and where appropriate family representatives. Such work should explore whether Sacred Pause™, Sovereign Brake, and Tiger .1x Key™ are experienced as supportive, obstructive, intuitive, or burdensome in different environments.

### 8.8.3 Calibration of High-Impact Thresholds

Another important area concerns the calibration of impact thresholds. The framework depends on distinguishing which outputs require stronger pause and authorization logic. Empirical studies should test how these classifications are determined and whether they are perceived as proportionate across diverse care contexts.

### 8.8.4 Validation and Refinement of Dignity Assessment

The Elder Dignity Score should be refined, operationalized, and validated through qualitative and mixed-method research. Dignity is unlikely ever to be fully capturable through a single metric, but structured tools can still help organizations assess whether technologies are preserving or undermining conditions of respectful care.

### 8.8.5 Comparative Studies

Future work should compare NAI 2.0™ against more conventional AI deployment models. Comparative research could examine differences in staff reliance, override behavior, trust calibration, incident management, privacy perception, and perceived dignity impacts. Such work would be especially valuable in demonstrating whether bounded governance offers measurable advantages over more permissive architectures.

### 8.8.6 Regulatory Translation and Standardization

Another promising area involves the translation of the framework into

standards, guidelines, or procurement criteria. Future research could explore how constitutional AI requirements for eldercare might be incorporated into institutional policy, assurance frameworks, audit templates, or pre-deployment review protocols.

### 8.8.7 Broader Application Beyond Eldercare

Although this thesis focused on eldercare, future studies may explore whether similar non-agentic constitutional approaches are appropriate in adjacent contexts such as disability support, mental health care, rehabilitation, or home-based long-term condition management. However, such extension should be done cautiously rather than assumed automatically.

## 8.9 Future Implementation and Development Directions

In addition to academic research, the thesis points toward several implementation-oriented next steps.

First, the framework would benefit from the development of a prototype deployment environment in which its core control logic can be tested safely. This should include simulation of care events, authorization pathways, pause triggers, override events, and audit review functions.

Second, implementation work should prioritize institutional co-design. Because eldercare is deeply contextual, organizations adopting such a framework would need involvement from caregivers, nurses, clinical leads, compliance officers, technologists, older adults, and families where appropriate. Co-design would help ensure that the controls remain meaningful without being disconnected from lived care realities.

Third, successful implementation will require training and cultural reinforcement. A system centered on pause and bounded authority cannot thrive in environments that treat speed as the sole marker of competence. Institutions would need to normalize reflective interruption as a legitimate feature of safe and dignified care.

Fourth, there is a need for procurement and governance templates that can help organizations evaluate whether AI vendors or internal systems meet bounded-authority requirements. In practice, many governance failures begin long before deployment, during selection and contracting.

Fifth, long-term development should explore how the framework can support continuous learning without authority expansion. One of the most difficult tensions in AI governance is allowing systems to improve without silently increasing their practical control. Future implementation models should ensure that adaptation remains bounded by the constitutional logic of the framework.

## 8.10 Final Reflections on Eldercare AI Governance

At a broader level, this thesis advances a normative argument about the future direction of AI in care. There is a powerful tendency in technological discourse to equate progress with increased automation, reduced friction, and ever more seamless delegation to machine systems. In many domains that assumption is already contested. In eldercare, it should be contested with particular seriousness.

Eldercare is a field in which efficiency matters, but it is not the only value. Safety matters, but so do privacy, explanation, recognition, and refusal. Prediction matters, but so does the right to contextual judgment. Systems that optimize speed or consistency may still fail if they weaken the human and relational conditions under which dignified care occurs.

This thesis therefore argues that the future of eldercare AI should not be guided by the question, How autonomous can the system become? Rather, it should be guided by the question, What forms of technological assistance are compatible with preserving human sovereignty over vulnerable care relationships? That change in orientation is, perhaps, the most important contribution of the study.

If that argument is accepted, then bounded AI is not a compromise with progress. It is a more mature understanding of what responsible progress requires.

## 8.11 Conclusion

This thesis set out to address a central challenge in contemporary eldercare innovation: the lack of a sufficiently robust governance architecture for AI systems operating in high-impact care environments. Through interdisciplinary analysis and design-oriented inquiry, it developed NAI 2.0™ as a hardware-enforced, non-agentic constitutional framework intended to preserve meaningful human sovereignty, dignity, and accountability.

The thesis demonstrated that current approaches to AI governance in healthcare and eldercare often remain too abstract or permissive to prevent authority drift. It argued that formal human oversight is not enough if systems are structured in ways that privilege machine-led progression. In response, it proposed a model in which pause, interruption, explicit authorization, bounded outputs, and traceable review are embedded into the architecture itself.

The findings of the study support a clear conclusion: NAI 2.0™ is a credible and substantial governance prototype for eldercare AI. It is conceptually coherent, architecturally plausible, ethically serious, and substantially aligned with the broader logic of emerging high-risk AI governance. At the same time, the thesis has remained cautious about its

limits. Real-world validation, usability testing, workflow calibration, and institutional implementation remain necessary before stronger empirical claims can be made.

Even so, the contribution is significant. The thesis has shown that eldercare AI need not be imagined primarily through the lens of expanding autonomy. It can instead be designed around restraint, sovereignty, dignity, and bounded support. In a field where vulnerable persons may be deeply affected by subtle shifts in authority, this is not a marginal design preference. It is a foundational governance imperative.

The final position of the thesis is therefore straightforward: AI can play a meaningful role in eldercare, but only if it remains constitutionally subordinate to accountable human judgment. NAI 2.0™ offers one serious and carefully reasoned path toward that goal.

# Chapter 9. Recommendations, Implementation Roadmap, and Strategic

Pathways

## 9.1 Introduction

While Chapter 8 concluded the core argument of this thesis, a further chapter is valuable in order to translate the study's conceptual, architectural, and governance findings into practical recommendations and forward-facing strategic action. If NAI 2.0™ is to matter beyond the level of doctoral analysis, it must be situated not only as a governance prototype, but also as a framework that can guide implementation decisions, institutional planning, regulatory interpretation, procurement logic, and future system development in eldercare.

This chapter therefore extends the thesis from evaluation to application. It does not introduce a new empirical data set, nor does it claim that implementation questions have been fully resolved. Rather, it identifies the most important actionable implications arising from the thesis and organizes them into a structured roadmap for stakeholders who may wish to adapt, pilot, assess, or govern bounded AI in eldercare settings.

The chapter proceeds from a central premise already established throughout the dissertation: eldercare AI should not be implemented as though technological capability alone justifies operational authority. Instead, implementation should be governed by constitutional restraint, meaningful human sovereignty, privacy-conscious design, and dignity-sensitive workflow logic. This means that recommendations cannot be limited to technical functionality. They must address institutional culture, role design, procurement criteria, audit structures, staff training, and the regulatory framing of AI-enabled care systems.

Accordingly, this chapter has six main purposes. First, it draws out the principal implementation implications of the thesis. Second, it presents practical recommendations for key stakeholder groups, including system designers, care providers, regulators, policymakers, and educators. Third, it proposes a phased implementation roadmap for NAI 2.0™ or similar bounded AI architectures. Fourth, it identifies governance and performance indicators that should be used during early deployment. Fifth, it considers strategic barriers to adoption and suggests mitigation approaches. Sixth, it offers a broader reflection on how eldercare systems may move toward a more sovereignty-preserving model of AI integration.

In doing so, this chapter functions as a bridge between doctoral theory and institutional action. It shows that the thesis's contribution is not only critical or conceptual, but also translational. The challenge now is not simply to understand why bounded AI governance is needed. It is to specify how such governance can be operationalized responsibly, proportionately, and credibly within the realities of eldercare practice.

## 9.2 Rationale for a Recommendations Chapter

The development of a governance framework is only one part of responsible innovation. A second and equally important task is to clarify the conditions under which that framework may be meaningfully adopted. This is especially true in eldercare, where implementation environments are often constrained by staffing pressure, fragmented governance structures, limited digital maturity, varying regulatory expectations, and understandable sensitivity to any intervention perceived as increasing burden or delaying response.

Without practical recommendations, there is a risk that the thesis's findings would remain at the level of normative aspiration. Yet one of the defining claims of this dissertation has been that governance must move beyond aspiration. If meaningful human sovereignty is to be preserved in AI-enabled eldercare, then institutions need more than ethical language. They need design criteria, procedural safeguards, decision thresholds, training expectations, audit requirements, and implementation sequencing.

A separate recommendations chapter is also justified because the thesis has shown that bounded AI governance requires coordinated action across multiple domains. No single stakeholder can deliver it alone. Designers can create pause mechanisms, but institutions determine whether pause is respected. Regulators can require oversight, but procurement teams determine whether systems are purchased with genuine boundedness. Care leaders can support dignity-sensitive use, but staff training shapes whether such use is feasible in practice. In short, successful governance depends on distributed responsibility.

This chapter therefore reflects the interdisciplinary character of the thesis itself. It takes seriously the proposition that eldercare AI is not merely a software matter. It is a matter of organizational design, professional norms, legal interpretability, infrastructural compatibility, and moral seriousness about the conditions under which vulnerable persons are cared for.

## 9.3 Core Implementation Principles Derived from the Thesis

Before presenting stakeholder-specific recommendations, it is helpful to restate the core principles that should guide any implementation of NAI 2.0™ or similar governance-by-design systems in eldercare.

### 9.3.1 Boundedness Before Capability Expansion

The first principle is that boundedness should precede capability expansion. Institutions should not begin with the question of how many functions AI can perform and then ask how to limit them later. Instead, they should define from the outset what categories of

action, escalation, inference, or workflow influence the system is not permitted to exercise independently.

### 9.3.2 Human Sovereignty Must Be Operationally Real

The second principle is that human sovereignty must be materially effective, not symbolic. This means that humans must have actual authority to pause, reject, override, or authorize consequential transitions at points where such intervention can still alter the outcome.

### 9.3.3 High-Impact Outputs Require Stronger Governance

The third principle is proportionality. Not all AI outputs require the same degree of friction. Low-impact administrative support functions may need lighter controls, while outputs affecting bodily welfare, surveillance intensity, restriction, escalation, or autonomy should trigger stronger pause and review logic.

### 9.3.4 Dignity Is a Design Constraint

The fourth principle is that dignity should be treated as a design condition, not a post hoc ethical slogan. Systems should be assessed in terms of whether they preserve explanation, refusal space, privacy proportionality, contextual judgment, and non-coercive care interaction.

### 9.3.5 Privacy Must Be Integrated Into Architecture

The fifth principle is that data governance cannot be delegated entirely to policy. Minimum necessary data use, role-limited access, clear purpose boundaries, retention discipline, and traceable auditability must be embedded into technical and operational structures.

### 9.3.6 Governance Must Anticipate Drift

The sixth principle is that implementation should assume the possibility of authority drift, normalization of override, and workflow shortcutting. Governance must therefore include periodic audit, cultural reinforcement, and mechanisms for identifying when the system is being used differently from how it was intended.

These principles synthesize the thesis's core findings and form the foundation for the recommendations that follow.

## 9.4 Recommendations for System Designers and Developers

The first group of recommendations is directed toward those involved in the design and development of AI-enabled eldercare systems.

### 9.4.1 Design for Non-Agentic Function From the Start

Developers should design eldercare AI systems explicitly as

non-agentic systems where high-impact outputs remain advisory, review-dependent, and incapable of self-executing consequential actions. This should be reflected not only in technical documentation, but also in system architecture, interface logic, and user permissions.

### 9.4.2 Separate Inference From Action

A crucial design recommendation is to maintain a strict distinction between AI inference generation and care action execution. Systems should be architected such that generating a recommendation, risk signal, or alert does not automatically create a consequential intervention pathway without explicit human review.

### 9.4.3 Build Non-Bypassable Pause Conditions

Where outputs affect high-impact care processes, developers should incorporate structured pause logic equivalent to Sacred Pause™. Such pauses should not be easily bypassed through convenience shortcuts, hidden defaults, or interface pressure toward immediate acceptance.

### 9.4.4 Enable Role-Specific Override and Authorization

Authorization systems should be role-sensitive. Not every user should have identical powers, and critical transitions should require appropriately credentialed human approval. This means that governance should be tied to actual care responsibilities and institutional accountability structures.

### 9.4.5 Reduce Interface-Induced Automation Bias

Interfaces should be designed to minimize over-trust. Recommendations should not be presented as if they are final decisions, nor should visual design imply unwarranted certainty. Where uncertainty exists, it should be made legible. Where contextual review is expected, the interface should prompt it directly.

### 9.4.6 Build Traceability Into the System Core

Every consequential recommendation, pause trigger, authorization action, override, and escalation should be logged in a way that supports audit, contestability, and post-event review. Traceability is essential not only for compliance, but for organizational learning and harm prevention.

### 9.4.7 Design for Minimum Necessary Data Use

Developers should adopt a data minimization approach. Inputs should be justified by purpose, and the system should avoid capturing unnecessary intimate behavioral or environmental data merely because such data could improve technical performance. In eldercare, more data is not always better governance.

### 9.4.8 Support Human Explanation

Outputs should be presented in a manner that allows care staff to explain to older persons or their representatives what the system identified, why attention is being drawn to a concern, and what remains subject to human review. Explanation supports both trust calibration and dignity preservation.

Overall, the major recommendation for designers is clear: eldercare AI should be built as constitutionally bounded infrastructure rather than as permissive software awaiting restraint through policy alone.

## 9.5 Recommendations for Care Providers and Eldercare Organizations

Care providers are the institutional actors most directly responsible for determining whether bounded AI governance becomes real in practice.

### 9.5.1 Adopt AI Only Within a Governance Framework

Organizations should avoid adopting eldercare AI as an isolated technological tool. Every deployment should be situated within a formal governance framework that defines intended use, prohibited uses, escalation rules, role responsibilities, override authority, audit pathways, and review procedures.

### 9.5.2 Create AI Use Policies Specific to Eldercare

Generic digital health policies are not enough. Institutions should develop eldercare-specific AI use protocols that address dignity, privacy, fluctuating capacity, family communication, contextual review, and the relational character of long-term care.

### 9.5.3 Preserve Human Review in Staffing Models

Human oversight cannot be meaningful if staffing structures make review impossible. Organizations should ensure that workflows provide enough time and role clarity for staff to engage with pause, review, and authorization requirements without treating them as obstacles to productivity.

### 9.5.4 Normalize Reflective Interruption

Leadership should actively reinforce the legitimacy of pause and challenge. Staff must feel authorized to question system outputs without being seen as resistant to innovation. If institutional culture rewards only speed and throughput, bounded AI governance will erode quickly.

### 9.5.5 Include Older Adults and Families in Governance Discussions

Where appropriate, older persons and family representatives should be informed about the role of AI in care pathways, especially where systems influence monitoring or escalation. This helps preserve trust and supports a more transparent care environment.

### 9.5.6 Monitor for Override Patterns and Drift

Organizations should not merely log override events; they should interpret them. Frequent override may indicate poor calibration, weak fit with real care context, or model limitations. Conversely, extremely rare override may signal over-trust or suppressed challenge culture.

### 9.5.7 Conduct Regular Governance Reviews

Institutions should establish periodic review meetings in which multidisciplinary teams examine incidents, near misses, escalation patterns, privacy concerns, and staff feedback related to AI use. Such reviews should be treated as part of routine governance, not exceptional crisis response.

### 9.5.8 Align Procurement With Bounded Authority Requirements

Procurement teams should require evidence that systems support pause, traceability, human authorization, privacy minimization, and role-based access control. Systems that optimize convenience by obscuring governance should not be treated as acceptable simply because they are technologically sophisticated.

The central recommendation for care organizations is that bounded governance must be institutionally protected, not merely technically available.

## 9.6 Recommendations for Regulators and Policymakers

Regulators and policymakers play a key role in shaping whether eldercare AI is governed seriously or superficially.

### 9.6.1 Move Beyond Formal Human-in-the-Loop Criteria

Regulatory guidance should distinguish clearly between nominal human involvement and meaningful human control. It is not enough for a system to include a person somewhere in the process. The relevant question is whether that person can actually influence or halt consequential outcomes.

### 9.6.2 Recognize Eldercare as a Distinct High-Sensitivity Context

Policy frameworks should explicitly recognize eldercare as a distinct AI governance environment. The combination of vulnerability, long-term dependency, privacy intimacy, and dignity-sensitive care justifies stronger guidance than generic health technology categories may provide.

### 9.6.3 Encourage Governance-by-Design Standards

Regulators should support standards that assess whether pause, interruptibility, explicit authorization, auditability, and purpose-bounded data use are structurally embedded in systems. Governance claims should be supported by design evidence rather than broad vendor declarations.

### 9.6.4 Promote Procurement Accountability

Policymakers should consider encouraging or requiring public and institutional procurement frameworks that examine bounded authority, role-specific control, privacy proportionality, and explainability in use. Procurement is one of the earliest and most influential governance points.

### 9.6.5 Require Clear Documentation of Intended Use and Non-Use

Systems should document not only their intended use but also their

non-use boundaries: what they are not designed to decide, authorize, or trigger independently. This would strengthen accountability and reduce post-deployment function creep.

### 9.6.6 Strengthen Audit Expectations

Regulatory expectations for high-impact AI should include robust logging of review, pause, override, and authorization events. Retrospective accountability matters because many governance failures become visible only when workflows are reconstructed.

### 9.6.7 Support Ethical Implementation Capacity

It is insufficient to regulate systems without supporting institutions in their implementation capacity. Policymakers should encourage training, governance infrastructure, interdisciplinary oversight teams, and practical implementation guidance for organizations with limited digital maturity.

The major policy message of the thesis is that eldercare AI regulation should evaluate control as exercised in practice, not merely as described in documentation.

## 9.7 Recommendations for Professional Education and Workforce Development

Even the strongest governance architecture can fail if the workforce is not prepared to use it appropriately.

### 9.7.1 Integrate AI Governance Into Eldercare Training

Training programmes for nurses, care staff, gerontology professionals, health informatics personnel, and administrators should include focused content on AI governance, authority drift, automation bias, and dignity-sensitive deployment.

### 9.7.2 Teach Staff to Interpret Rather Than Obey Outputs

Educational efforts should emphasize that AI outputs are inputs to judgment, not replacements for it. Staff should be trained to assess context, uncertainty, and relevance rather than treating machine recommendations as inherently superior.

### 9.7.3 Prepare Leaders for Governance Responsibility

Operational leaders need training not only in technology adoption but also in governance stewardship. This includes understanding audit patterns, escalation pathways, staff burden, and the cultural conditions required for meaningful pause and override.

### 9.7.4 Develop Scenario-Based Simulation

Simulation exercises should be used to train staff in handling pause events, override decisions, and high-impact recommendations. Such simulations can help normalize reflective engagement and reveal workflow weaknesses before live deployment.

### 9.7.5 Include Dignity and Communication Skills

Education should address the interpersonal dimensions of AI-enabled care. Staff must be able to explain system involvement in ways that preserve respect, reduce fear, and maintain relational trust with older adults and families.

The main educational implication is that bounded AI governance is a professional competency issue as much as a technical one.

## 9.8 Recommendations for Future Researchers

Researchers have an important role in extending the work begun in this thesis.

### 9.8.1 Prioritize Real-World Pilot Studies

Future research should test bounded AI architectures in real eldercare environments through carefully governed pilot studies. These studies should examine feasibility, staff behavior, override patterns, trust calibration, and perceived impact on dignity and safety.

### 9.8.2 Develop Mixed-Method Evaluation Models

Because eldercare AI raises operational and ethical questions simultaneously, future studies should combine workflow metrics, audit data, qualitative interviews, ethnographic observation, and patient-family perspectives where appropriate.

### 9.8.3 Refine Measures of Dignity-Related Impact

Researchers should continue developing structured but cautious tools for assessing dignity-related effects, including perceived intrusiveness, loss of refusal space, depersonalization risk, and relational quality of care under AI-supported workflows.

### 9.8.4 Compare Bounded and Permissive Architectures

Comparative studies are needed to examine whether bounded governance models produce different outcomes from more permissive AI designs in terms of trust, safety, burden, error response, and human oversight quality.

### 9.8.5 Study Governance Drift Over Time

Longitudinal research should investigate how AI systems change in practice after deployment. One of the greatest risks is that bounded systems become less bounded through routine workarounds or cultural normalization. Research should track such drift directly.

### 9.8.6 Explore Cross-Jurisdictional Application

Further work should examine how NAI 2.0™-like frameworks could be adapted to different legal, cultural, and health system contexts. Eldercare norms vary internationally, and bounded AI governance will need contextual calibration.

Research should therefore move beyond capability testing and engage more deeply with authority design, institutional behavior, and dignity-sensitive evaluation.

## 9.9 A Phased Implementation Roadmap for NAI 2.0™

To support practical translation, this section proposes a phased roadmap for implementation.

### 9.9.1 Phase 1: Governance Readiness Assessment

Before deployment, institutions should conduct a readiness review assessing:

- current digital infrastructure,
- staffing capacity for review and authorization,
- privacy governance maturity,
- existing incident reporting systems,
- leadership support for pause-based workflows,
- and compatibility with bounded authority principles.

If an institution is unable to support meaningful human review, deployment should be delayed rather than rushed.

### 9.9.2 Phase 2: Scope Definition and Intended Use Framing

Organizations should define:

- the exact eldercare use cases,
- risk categories involved,
- outputs considered high impact,
- prohibited autonomous actions,
- required human roles for review,
- and acceptable decision boundaries.

This phase is crucial for preventing ambiguity and future function creep.

### 9.9.3 Phase 3: Technical Configuration and Safeguard Embedding

At this stage, the system should be configured with:

— pause thresholds,

— authorization roles,

— audit logging,

— data minimization settings,

— access permissions,

— and hardware-linked enforcement where applicable.

Testing should focus on whether the controls genuinely work under realistic workflow conditions.

### 9.9.4 Phase 4: Workforce Training and Simulation

Before live use, staff should receive structured training on:

— the role of the system,

— what the system may and may not do,

— how Sacred Pause™ functions,

— how Sovereign Brake is exercised,

— when Tiger .1x Key™ authorization is required,

— and how to document concerns or override actions.

Simulation should include routine, ambiguous, and high-pressure scenarios.

### 9.9.5 Phase 5: Limited Pilot Deployment

Initial deployment should be narrow in scope, involving carefully chosen environments and strong oversight. During this phase, organizations should monitor:

— number of pause events,

— authorization patterns,

— override frequency,

— user concerns,

— privacy incidents,

— and workflow disruption signals.

Pilot deployment should be evaluative, not merely demonstrative.

### 9.9.6 Phase 6: Review, Recalibration, and Governance Correction

Following pilot use, institutions should review whether the system:

— preserved meaningful human control,

— triggered appropriate levels of friction,

— aligned with intended use,

— created unexpected burden,

— or showed signs of authority drift.

Thresholds, permissions, and workflow integration should be recalibrated before any scale-up.

### 9.9.7 Phase 7: Controlled Scale-Up

Only after successful evaluation should broader deployment be considered. Even then, expansion should be gradual and accompanied by periodic governance review, refresher training, and independent audit where feasible.

This phased roadmap reflects the thesis's central proposition that

bounded AI deployment must be earned through governance readiness, not assumed through technical enthusiasm.

## 9.10 Suggested Governance and Monitoring Indicators

To operationalize oversight, institutions should monitor clear indicators during deployment.

### 9.10.1 Control and Review Indicators

— frequency of Sacred Pause™ activation,

— average time spent in review,

— percentage of high-impact outputs receiving documented human assessment,

— number and pattern of override decisions,

— and proportion of actions requiring Tiger .1x Key™ authorization.

### 9.10.2 Workflow and Feasibility Indicators

— staff-reported usability,

— delay burden in urgent but non-emergency pathways,

— training completion rates,

— workarounds or bypass attempts,

— and system downtime or interruption incidents.

### 9.10.3 Dignity and Privacy Indicators

— complaints related to intrusiveness or loss of privacy,

— family or resident concerns about surveillance,

— documented explanation practices,

— use of minimum necessary data,

— and incidents of access beyond role necessity.

### 9.10.4 Governance and Accountability Indicators

— audit completeness,

— review meeting frequency,

— action taken following identified drift,

— policy deviations,

— and proportion of AI-related incidents receiving formal governance analysis.

These indicators should not be treated as purely administrative metrics. They are signals of whether the system is remaining faithful to its constitutional purpose.

## 9.11 Anticipated Barriers to Adoption

A serious implementation strategy must also anticipate resistance and difficulty.

### 9.11.1 Workflow Burden Concerns

Staff may perceive pause and authorization mechanisms as slowing care processes. This concern is legitimate and should not be dismissed. It reinforces the need for careful threshold calibration and targeted rather than indiscriminate friction.

### 9.11.2 Cultural Preference for Seamlessness

Many innovation narratives privilege frictionless systems. As a result, bounded governance may initially appear regressive or inefficient. Institutions will need to communicate clearly that in high-impact eldercare, thoughtful friction is a safety and dignity feature, not simply a technical limitation.

### 9.11.3 Resource Constraints

Some organizations may lack the digital infrastructure or staffing model needed to support meaningful review. In such cases, bounded AI governance may appear difficult to implement. Yet this does not justify weaker governance; it indicates that institutional readiness must precede deployment.

### 9.11.4 Vendor Resistance

Vendors may resist governance requirements that reduce claims of seamless automation or increase development burden. Procurement criteria and regulatory expectations will be critical in shifting incentives toward bounded design.

### 9.11.5 Drift Through Routine Practice

Even well-designed systems may be weakened over time by routine bypass, shortcutting, or cultural habituation. This is why periodic audit and reinforcement are essential.

Recognizing these barriers does not undermine the thesis. It strengthens it by acknowledging that governance success depends on real organizational conditions.

## 9.12 Strategies for Overcoming Implementation Barriers

Several strategies can improve the chances of successful bounded AI adoption.

### 9.12.1 Frame Governance as Care Protection

Institutions should explain that pause, authorization, and review are not obstacles to innovation but mechanisms for protecting residents, staff, and organizational accountability.

### 9.12.2 Use Risk-Stratified Controls

Not every output should trigger identical burden. By applying stronger controls only where justified, organizations can preserve workflow practicality while maintaining constitutional restraint where it matters most.

### 9.12.3 Engage Staff Early

Frontline staff should be involved in design adaptation, pilot planning, and evaluation. Systems imposed without participatory engagement are more likely to produce workarounds and mistrust.

### 9.12.4 Build Leadership Accountability

Senior leaders should own governance outcomes, not delegate them entirely to technical teams. Cultural legitimacy for bounded AI begins with leadership.

### 9.12.5 Make Audit Actionable

Audit data should feed directly into governance corrections, retraining, and system refinement. Logging without responsive action produces little value.

### 9.12.6 Maintain Transparency With Residents and Families

Open communication about the role and limits of AI can help preserve trust and reduce fear that technology is silently replacing human care.

## 9.13 Strategic Vision Beyond the Thesis

The broader strategic implication of this chapter is that NAI 2.0™ should not be understood merely as a single framework, but as a model for a different direction of AI innovation in care. Much current technological discourse moves toward increasing automation, predictive dependency, and seamless integration. This thesis has proposed another route: one in which intelligence remains useful but constitutionally subordinate.

If this direction is taken seriously, several wider shifts may follow. Procurement may begin to reward bounded systems rather than merely high-performing ones. Regulators may scrutinize whether human oversight is operationally genuine. Care organizations may begin to see reflective interruption as compatible with quality rather than opposed to it. Designers may start to ask not only whether a system works, but whether it works without displacing the moral center of care.

In that sense, the strategic future envisioned here is not anti-innovation. It is disciplinarily mature innovation. It accepts that in eldercare, progress cannot be measured only by automation depth or response speed. It must also be measured by whether older adults remain protected against silent transfers of authority from human care relationships to machine-driven workflow.

## 9.14 Chapter Summary

This chapter has translated the thesis's conceptual and evaluative findings into a structured set of recommendations and strategic directions for implementation. It has argued that the success of bounded eldercare AI depends on coordinated action across design, institutional governance, regulation, education, and research. It has proposed core implementation principles, stakeholder-specific recommendations, a phased roadmap for deployment, practical governance indicators, and strategies for addressing likely barriers to adoption.

The central message of the chapter is consistent with the overall argument of the dissertation: eldercare AI should be introduced only within a governance architecture capable of preserving meaningful human sovereignty, privacy-conscious operation, and dignity-sensitive care. Bounded authority is not an abstract ideal. It is something that must be procured, designed, trained for, audited, and culturally reinforced.

This chapter therefore extends the contribution of the thesis from doctoral analysis toward practical application. It shows that NAI 2.0™ is not only a theoretical response to the risks of authority drift, but also a framework capable of informing real institutional strategy.

**Concluding Note to Chapter 9**

The final task of this thesis has been to move from argument to direction. Having established the need for a constitutional, non-agentic AI governance framework in eldercare, this chapter has shown how such a framework might begin to shape actual policy, design, and deployment choices. The recommendations offered here do not eliminate the need for future testing, adaptation, or empirical validation. But they do make one thing clear: bounded AI governance is not impractical by definition. It becomes impractical only when institutions attempt to adopt AI without taking governance seriously enough.

NAI 2.0™ was developed in this thesis as a serious alternative to weakly governed autonomy-seeking systems in eldercare. This chapter affirms that the alternative is actionable. If future stakeholders choose to build, regulate, procure, and deploy eldercare AI through the lens of sovereignty, dignity, and constitutional restraint, then the work begun in this dissertation can become more than a theoretical intervention. It can become part of a wider shift in how care technologies are imagined and governed.

# References


Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

Beauchamp, T. L., & Childress, J. F. (2019). Principles of biomedical ethics (8th ed.). Oxford University Press.

Brown, B., et al. (2023). Micro-friction interventions and System 1 cognition in clinical AI. Journal of Medical AI, 4(2), 112–128.

Coeckelbergh, M. (2020). AI ethics. MIT Press.

European Parliament. (2024). EU Artificial Intelligence Act. Official Journal of the European Union.

Floridi, L., et al. (2018). AI4People — An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias — empirical results. International Journal of Medical Informatics, 83(5), 368–375.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105.

Koh, E. W. K. (2026). ABC+2S+H Guardian Framework™ [P-001]. IPOS SG020603109STW. ACRA T260229801.

Koh, E. W. K. (2026). WD070–WD073: Constitutional Drift Governance Patents. NLB R260219-005.

Koh, E. W. K. (2026). WD113–WD117: Eldercare Deployment Patent Series. NLB R260219-005.

Koh, E. W. K. (2026). WM001–WM009: Wearable Medical Patent Series. NLB R260219-005.

Koh, E. W. K. (2026). WISL No. 01–25: Constitutional White Paper Canon. NLB R260302-007.

Low, S. E. (2026, April 2). RE: Non-Agentic Eldercare AI — Academic-Non-Commercial Collaboration. NTU LKCMedicine.

Manzeschke, A., et al. (2023). Ethical frameworks for assistive AI in elder care. Gerontechnology, 22(1), 1–14.

Ministry of Health Singapore. (2021). Artificial Intelligence in Healthcare Guidelines (AIHGle). MOH Holdings / HSA / IHiS.

NTU ARISE. (2026). Ageing Research Institute for Society and Education — Research Domains. Nanyang Technological University.

Pew, R. W., & Mavor, A. S. (1998). Human-system integration in the system development process. National Academy Press.

Rasmussen, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols. IEEE Transactions on Systems, Man, and Cybernetics, 13(3), 257–266.

Regulation (EU) 2016/679. (2016). General Data Protection Regulation. European Parliament.

Schwartz, M. D. (2019). Medical ethics: A case-based approach (3rd ed.). Elsevier.

Shneiderman, B. (2022). Human-centered AI. Oxford University Press.

Singapore Computer Society. (2026). SCS AI Ethics & Governance Body of Knowledge Version 2.0.

Sokol, K., et al. (2025). Reflective anchoring and AI-assisted clinical decisions. BMJ Health and Care Informatics, 12(1).

Stoumpos, A. I., Kitsios, F., & Talias, M. A. (2023). Digital transformation in healthcare: Technology acceptance and its applications. International Journal of Environmental Research and Public Health, 20(4), 3407.

Vered, M., Shani, G., & Zahavi, H. (2023). Explanations and enforced deliberation for reduced automation bias. Computers in Human Behavior, 148, 107929.

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and AI systems. Philosophical Transactions of the Royal Society A, 376(2133).

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO.

Zerilli, J., et al. (2019). Algorithmic decision-making and the control problem. Minds and Machines, 29(4), 555–578.