A Philosophical Approach to Judicial AI: The Role of Judgment in The Decision-Making Process

Ilaria Iannuccilli[1,*]

[1] University of Rome Tor Vergata, Via Columbia 1, Rome, 00133, Italy

Abstract

This paper examines the relationship between artificial intelligence and human decision-making processes through a philosophical lens. It evaluates how AI systems, while beneficial for enhancing moral reasoning and reducing biases, risk causing "deresponsibilization" - where users become passive executors rather than responsible agents. Drawing on Hannah Arendt's theory of judgment, L.A. Paul's concept of "transformative experience," Günther Anders' warnings about technology delegation, and Shannon Vallor's concerns about moral deskilling, the paper argues for maintaining human agency in AI-assisted ethical decisions. It specifically analyzes Judicial AI protocols that present multiple interpretations to users, preserving their reflective capacity and responsibility. Lastly, a simulated scenario highlights that while AI systems can aid in complex ethical dilemmas, humans must continue cultivating moral virtues through practice to avoid surrendering their moral agency and missing opportunities for personal transformation through independent decision-making.

Keywords

Judicial AI, Judgment, Transformative Experience, Decision-Making Process

1. Introduction

Judgment is a key human faculty within the decision-making process that involves critical thinking and responsibility. We exercise this faculty especially when ethical and moral dilemmas need to be faced. Nowadays, more AI technologies are developed in order to be able to support humans in solving these kind of dilemmas. Although these systems are useful tools, able to enhance human reasoning and help humans to make better moral choices and avoid biases, they could also induce an over-reliance on AI that can result in a loss of skills and responsibility. It is worth recalling the concept of deresponisibilization that "reflects the risk that users may start to perceive themselves as mere executors of the system's decisions rather than active, responsible agents, which are accountable for the final decision" [1]. This paper focuses on Judicial AI with a philosophical approach based on the theories of Arendt, Paul, Anders and Vallor. It shows how a philosophical framework based upon the aforementioned philosophical theories can implement an AI system during the decision-making process. It examines how Arendt's concept of judgment plays a crucial role in human decision-making processes and the importance of Anders machine delegation warning. It further explores Paul's theory of "transformative experience," which describes the profound changes individuals undergo through decision-making process, alongside Vallor's framework addressing the risk of moral deskilling when humans delegate ethical decisions to AI systems. This paper also features a simulation of a healthcare resource allocation scenario that demonstrates how philosophical frameworks can implement Judical AI protocols to preserve human moral agency in critical decision-making situations.

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy

[*] Corresponding author.

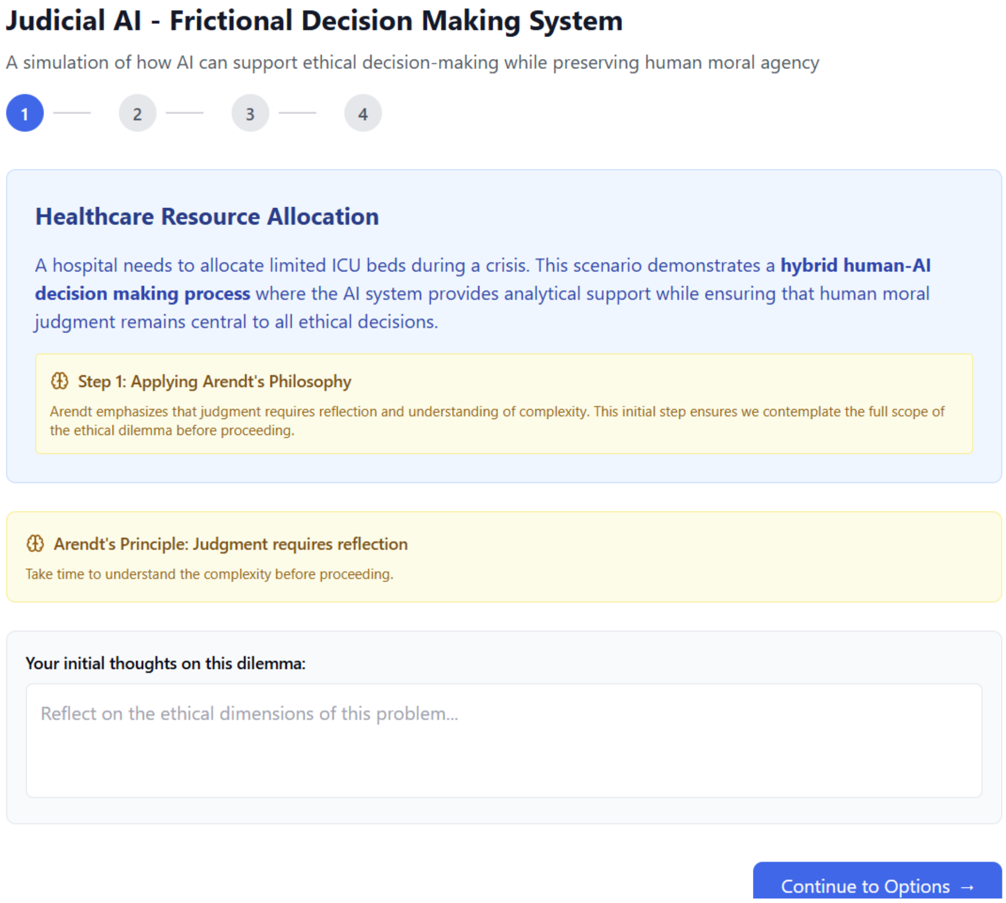
i.iannuccilli@gmail.com (I. Iannuccilli)
0009-0005-6864-021X (I. Iannuccilli)

2. Arendt's Concept of Judgment and Judicial AI

Judicial AI is a protocol related to Frictional AI, in which "the AI system provides arguments and explanations that support multiple, conflicting decisions or interpretations" [2]. In this case, I address to the experiment lead by Fregosi and Cabitza that investigates "the textual generative setting in Judicial protocol for sentence classification" [1]. This experiment shows that the use of a judicial protocol "in human-AI interaction is expected to have a significant impact on the quality of decisions made by users, in particular on perceived agency and control over their choices" [1]. The possibility to introduce a moment of reflection, while all the available options are evaluated, is necessary to make the choice in a fully responsible way. To be accountable of our choices means to be responsible: in Arendt thought human responsibility and thinking are necessary in order to exercise judgement which is what drives human action. Arendt's theory of judgment, one of her most relevant elaboration, is a useful tool because it deals with the fact that a person always finds themselves facing particular and contingent situations, and to navigate them, they do not have general criteria already prepared. This is especially true in the ethical field, where general principles are often not sufficient to decide regarding concrete cases [3]. Arendt's notion of Socratic dialogue calls us to cultivate an inner dialogue with ourselves in order to deal with otherness and to build our personal morality. In this sense, humans can be fully conscious of their choices because they can take responsibility for their decisions and actions. It is evident that in a decision-making process involving an AI system, the implementation of a protocol such as that of Judicial AI is preferable, as it facilitates greater opportunities for the exercise of judgment by asserting human agency.

3. Transformative Experience and The Risk of Moral Deskilling

The application of AI systems, such as Artificial Moral Agents (AMAs) or Decision Support Systems (DDS), in decision-making processes is becoming increasingly common in a wide variety of fields. As Myers and Everett [4] state "AI systems are approaching a level of complexity that progressively requires them to embody artificial morality". Although the use of an AI system that shows a certain degree of morality is now desirable, because it can function as a tool to help people resolve more complex moral issues, the risk for users might result in a moral deskilling and in a loss of agency. Vallor states that "moral skills are intrinsically valuable" [5] and humans must not lose these skills "even if intelligent machines could somehow direct all human interactions to produce the most just, harmonious, and compassionate outcomes possible" [5] because "we would be diminished as creatures were we utterly helpless to act justly and compassionately without their assistance" [5]. In this sense, it is necessary to continue cultivating our virtues and, therefore, our human skills in order to be held as responsible agents. Virtues, indeed, "have to be cultivated through practice; we acquire virtues by doing; not by mere wishing" [6]. In this regard, it is useful to report the analysis of the philosopher Anders who, during the 1950', highlighted a tendency among people to delegate their choices to a machine. He states that human beings are ready to recognize an Electric Brain as a surrogate for their conscience [7]. Out of this situation it comes a sense of humiliation for human beings that derives from the loss of responsibility for their own choices. Indeed, when a person makes a choice or a decision this determines a change in their life. Paul states that "making the choice in the right way is a way of taking charge of your own life. Choices involve responsibility, and to choose responsibly, you need to assess how your choice will affect the world and others in your life" [8]. Therefore, during the decision-making process, we are responsible of our choices not only towards ourselves but also towards others, and moreover, the changes that result from it may transform us in a radical way. This is what Paul calls "transformative experience"[8]: it is a radical change that modifies the person that has made the choice. Transformative experiences can be epistemically and/or personally transformative; in the first case the experience transforms what we know and understand while in the second case it transforms our personal and particular preferences. This theory, therefore, highlights the risks of an over-reliance on AI systems in the decision-making process: not only the loss of agency but also the loss of opportunities



Judicial AI - Frictional Decision Making System

A simulation of how AI can support ethical decision-making while preserving human moral agency

Healthcare Resource Allocation

A hospital needs to allocate limited ICU beds during a crisis. This scenario demonstrates a **hybrid human-AI decision making process** where the AI system provides analytical support while ensuring that human moral judgment remains central to all ethical decisions.

Step 1: Applying Arendt's Philosophy

Arendt emphasizes that judgment requires reflection and understanding of complexity. This initial step ensures we contemplate the full scope of the ethical dilemma before proceeding.

Arendt's Principle: Judgment requires reflection

Take time to understand the complexity before proceeding.

Your initial thoughts on this dilemma:

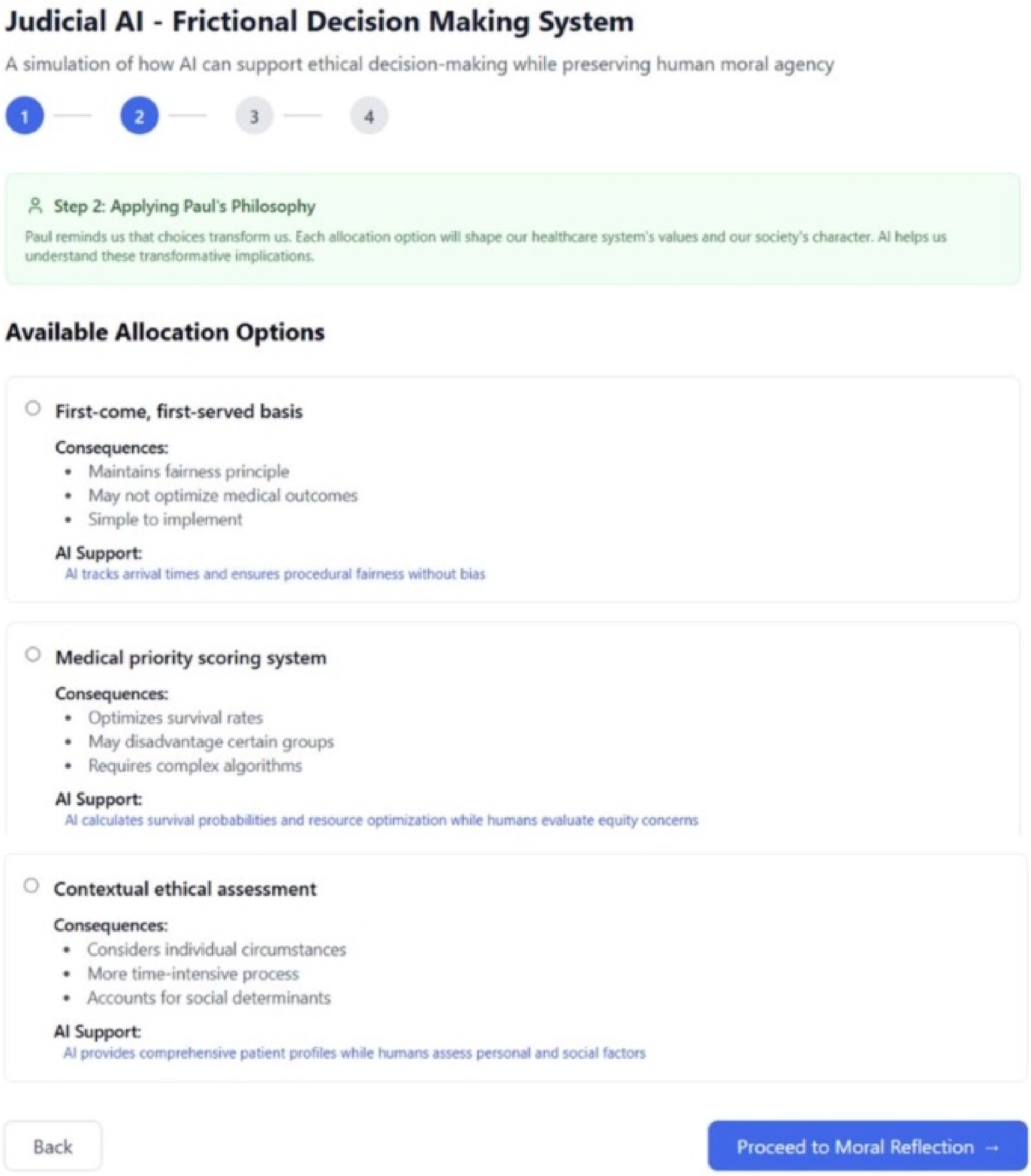
Reflect on the ethical dimensions of this problem...

Figure 1: Section 1: Judgment

to gain new knowledge through new experiences resulting from the faculty of judgment of each person.

4. A Philosophical Implementation

The simulation of how a cooperation between human and AI could be helpful during the decision-making process was built on the case of healthcare resource allocation. With the support of the AI program Claude Sonnet 4 (Anthropic), the simulation demonstrates how a Judicial AI protocol with a philosophical implementation can support humans during decision-making processes of complex cases. In the first section (figure1), the interface shows the problem (an hospital must allocate limited icu beds during an emergency) and presents the first philosophical theory that will guide the reflection. The user is asked to provide her initial thoughts on the problem. This is the first moment of reflection where the user became aware of the impact of her decision. The user is asked to assess the problem and understand its complexity by exercising her judgment and to write her initial thoughts in the open question. The second section (figure 2), is guided by Paul's philosophical insight of transformative experience: every decision we make has an impact not only to the immediate problem but also to ourselves as individual in the society. The three options are related to fairness, efficiency and human value and each one of them presents their consequences and AI support. While the AI support is useful, the role of human agency and judgment remains crucial. The third section (Figure 3) implements Vallor's theory of virtue cultivation, necessary in order to avoid the risk of over reliance to AI recommendations. The critical questions for reflection help the user to reflect deeper and to have an active engagement with the ethical principles. The user has to provide moral reasoning of her choice while the mandatory reflection period (the timer duration is illustrative) operates as a "moral friction" because the user can't skip the countdown but instead has to wait its end before sending the reply. The moral 'work' of reflection,



Judicial AI - Frictional Decision Making System

A simulation of how AI can support ethical decision-making while preserving human moral agency

Step 2: Applying Paul's Philosophy

Paul reminds us that choices transform us. Each allocation option will shape our healthcare system's values and our society's character. AI helps us understand these transformative implications.

Available Allocation Options

○ **First-come, first-served basis**

Consequences:
- Maintains fairness principle
- May not optimize medical outcomes
- Simple to implement

AI Support:
AI tracks arrival times and ensures procedural fairness without bias

○ **Medical priority scoring system**

Consequences:
- Optimizes survival rates
- May disadvantage certain groups
- Requires complex algorithms

AI Support:
AI calculates survival probabilities and resource optimization while humans evaluate equity concerns

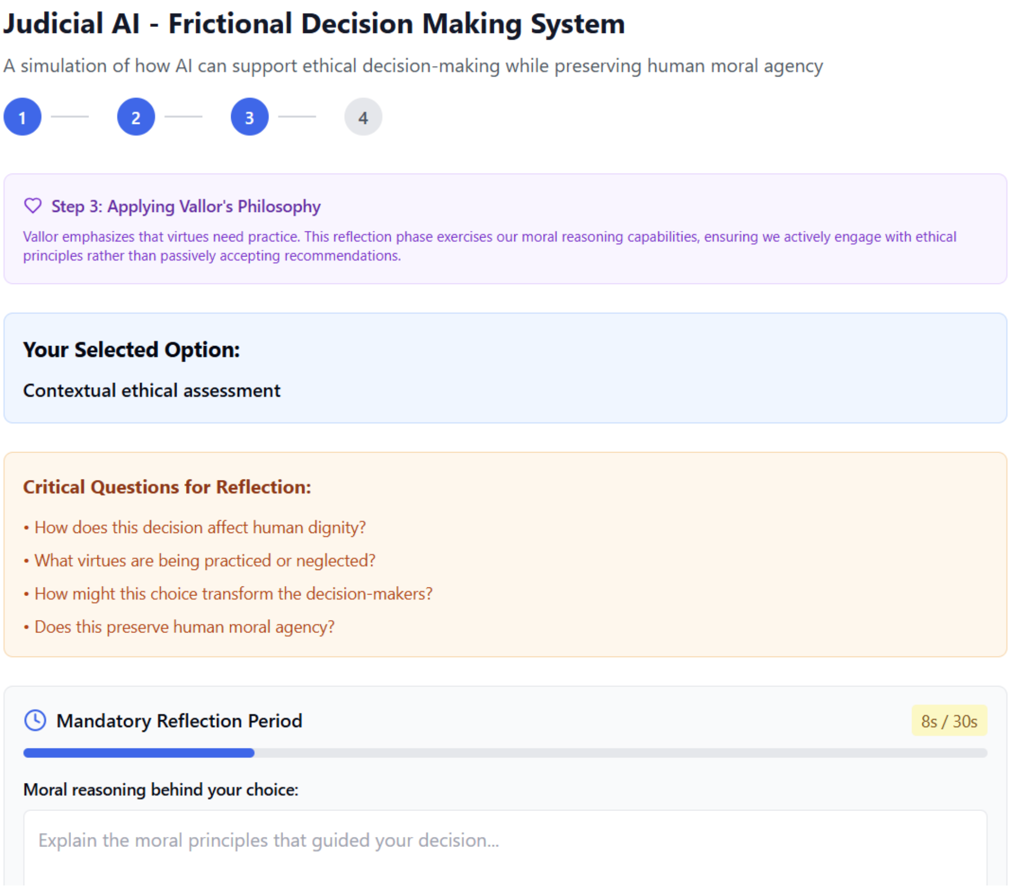
○ **Contextual ethical assessment**

Consequences:
- Considers individual circumstances
- More time-intensive process
- Accounts for social determinants

AI Support:
AI provides comprehensive patient profiles while humans assess personal and social factors

Figure 2: Section 2: Transformative Experience

reasoning, and decision remains with humans. The fourth section (Figure **??**) applies Anders' theory and warnings about preserving human conscience within the technological systems. This section represents a checkpoint where the user evaluates her responses starting with the critical question built on Anders' insight to detect whether the user was operating as a genuine moral agent or as a follower of AI recommendations. After the decision summary, the user can decide whether to start over the decision-making progress or to confirm the decision. The user is called to take responsibility of her decision: in this sense the kind of diffused responsibility that would result from human-machine collaboration is prevented. The simulation illustrates that human oversight, human responsibility, and human moral agency can be protected through the process. The system introduces moments of "friction" or "moral friction" such as structured pauses, reflection requirements, and philosophical prompts that slow down the decision-making process to ensure thoughtful ethical engagement and



Judicial AI - Frictional Decision Making System

A simulation of how AI can support ethical decision-making while preserving human moral agency

Step 3: Applying Vallor's Philosophy

Vallor emphasizes that virtues need practice. This reflection phase exercises our moral reasoning capabilities, ensuring we actively engage with ethical principles rather than passively accepting recommendations.

Your Selected Option:

Contextual ethical assessment

Critical Questions for Reflection:

- How does this decision affect human dignity?
- What virtues are being practiced or neglected?
- How might this choice transform the decision-makers?
- Does this preserve human moral agency?

Mandatory Reflection Period

8s / 30s

Moral reasoning behind your choice:

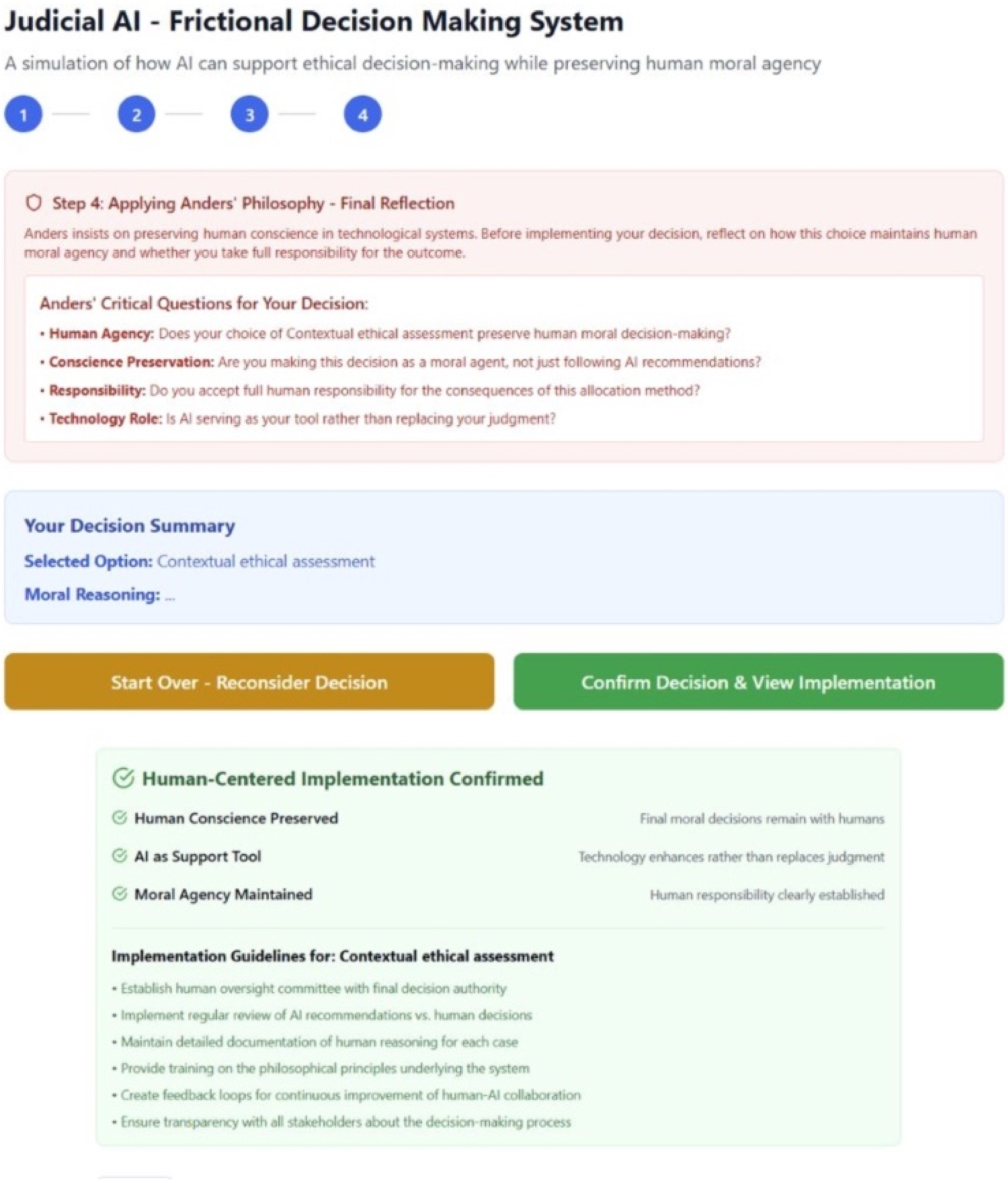
Explain the moral principles that guided your decision...

Figure 3: Section 3: Virtue Cultivation

judgment exercise. The system includes mandatory waiting periods, requires written moral reasoning, and offers users the ability to reconsider their decisions entirely.The simulation provides a template for how AI systems can be designed to enhance rather than replace human moral judgment with philosophical theories implementation. Each section creates a comprehensive framework for ethical decision-making in high-stakes scenarios.

5. Conclusion

This paper highlights the critical balance between AI assistance and human responsibility in ethical decision-making. While AI systems with moral capabilities can support humans in resolving complex ethical dilemmas, we must remain vigilant against the risk of over-reliance that leads to moral deskilling and the delegation of responsibility. Arendt’s theory of judgment emphasizes the necessity of human reflection in contextual ethical situations, while Paul’s concept of "transformative experience" reminds us that we should make decisions consciously because they transform us epistemically and/or personally. Anders’ prescient warning about humans delegating responsibility to machines underscores the humiliation that comes from surrendering our moral agency. To preserve what makes us human, Vallor emphasizes that we must continue to cultivate moral virtues through practice and maintain our role as responsible agents, even as we benefit from AI support. The simulation of health allocation resources case demonstrates that the implementation of protocols like Judicial AI can help maintain this balance by providing options that still require human judgment, thus preserving our essential capacity for responsible decision-making.



1
2
3
4
Start Over - Reconsider Decision
Confirm Decision & View Implementation

Judicial AI - Frictional Decision Making System

A simulation of how AI can support ethical decision-making while preserving human moral agency

Step 4: Applying Anders' Philosophy - Final Reflection

Anders insists on preserving human conscience in technological systems. Before implementing your decision, reflect on how this choice maintains human moral agency and whether you take full responsibility for the outcome.

Anders' Critical Questions for Your Decision:

- **Human Agency:** Does your choice of Contextual ethical assessment preserve human moral decision-making?
- **Conscience Preservation:** Are you making this decision as a moral agent, not just following AI recommendations?
- **Responsibility:** Do you accept full human responsibility for the consequences of this allocation method?
- **Technology Role:** Is AI serving as your tool rather than replacing your judgment?

Your Decision Summary

Selected Option: Contextual ethical assessment

Moral Reasoning: ...

Human-Centered Implementation Confirmed

✓ Human Conscience Preserved — Final moral decisions remain with humans

✓ AI as Support Tool — Technology enhances rather than replaces judgment

✓ Moral Agency Maintained — Human responsibility clearly established

Implementation Guidelines for: Contextual ethical assessment

- Establish human oversight committee with final decision authority
- Implement regular review of AI recommendations vs. human decisions
- Maintain detailed documentation of human reasoning for each case
- Provide training on the philosophical principles underlying the system
- Create feedback loops for continuous improvement of human-AI collaboration
- Ensure transparency with all stakeholders about the decision-making process

Figure 4: Section 4: Responsibility

6. Declaration on Generative AI

Figures 1, 2, 3, 4 were generated by the AI tool *Claude Sonnet 4 Anthropic*

References

[1] C. Fregosi, F. Cabitza, A frictional design approach: Towards judicial ai and its possible applications (2024).

[2] C. Natali, et al., Per aspera ad astra, or flourishing via friction: Stimulating cognitive activation by

design through frictional decision support systems, in: CEUR workshop proceedings, volume 3481, CEUR-WS, 2023, pp. 15–19.

[3] H. Arendt, The life of the mind, 1977.

[4] S. Myers, J. A. Everett, People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors, Cognition 256 (2025) 106028.

[5] S. Vallor, Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character, Philosophy & Technology 28 (2015) 107–124.

[6] S. Vallor, The AI mirror: How to reclaim our humanity in an age of machine thinking, Oxford University Press, 2024.

[7] P. Anders, L'uomo è antiquato.: Volume I, Considerazioni sull'anima nell'epoca della Seconda rivoluzione industriale, volume Volume I, Bollati Boringhieri, 2007.

[8] J. Schwenkler, E. Lambert, Becoming someone new: Essays on transformative experience, choice, and change (2020).