Ethical AI Triplet: A Framework for Stress-Testing Fairness in Digital Twins in Healthcare

Daniel Ioan Chis*, Ioan Dumitrache**

**National University of Science and Technology Politehnica Bucuresti, (chisdanielioan@gmail.com)*
***Romanian Academy, Bucuresti, Romania (ioan.dumitrache@acad.ro)*

Abstract: With "Digital Twins" tools integrating more Artificial Intelligence agents, those become powerful tools for optimizing domains like public health. Ensuring their ethical alignment with core medical rules (e.g. empathy, equity, justice etc.) is paramount. This paper proposes a counter framework to balance them: the Ethical AI Triplet. Our novel architecture acts as an automated "ethics committee" that validates policies proposed by an optimization AI. The "Triplet" has three specialized agents: a Population Generator that creates diverse societies, a Simulation Engine that runs proposed policies against the societies, and an Ethical Auditor that evaluates the outcomes against metrics of fairness and harm. Unlike traditional AI governance, which relies on post-hoc audits and manual reviews, the Ethical AI Triplet showcases an automated, adversarial framework designed to stress test policies for ethical robustness before deployment, identifying in time harmful possibilities that existing frameworks may miss.

Keywords: Ethical AI, Digital Twins, AI Governance, Complex Systems, Policy Validation, AI Safety, AI Red Teaming, eHealth Ethics

1. INTRODUCTION: THE NEED FOR ETHICAL VALIDATION

As industries take a leap forward and start to integrate AI agents into their infrastructure and workflows, new strategies need to be put into place in order to validate that those new technologies are truly beneficial in their use, especially in domains that directly affect humans, like healthcare.

As Artificial Intelligence (AI) is integrated into this space more and more, it promises a shift of paradigm, opening the opportunities to accelerate drug discovery, create deeper and broader personalized treatments, new enhancements to diagnostic accuracy, all on an unprecedented scale (Topol, 2019).

A new frontier in this domain is the creation and development of **Digital Twins** for public health and for medical care.

A digital twin represents a dynamic, virtual representation of a real-world system that is continuously updated with data from the physical counterpart (Grieves and Vickers, 2017).

The application of those new technologies in healthcare has seen a rapid advancement, being used in multiple areas from individual to population level needs.

Recent studies showcase their usefulness in creating patient specific models for personalized oncology treatments (Sun et al., 2023) or ophthalmology treatment (Iliuta et al., 2024), real-time management of chronic diseases like diabetes (Martin et al., 2023) and even being able to model entire urban environments to test public health emergency responses (Liu et al., 2024).

The usefulness of the twins is unlocked at a population level when an embedded AI Policy Optimizer uses different situations to test thousands of potential health strategies and discover novel, high-leverage opportunities.

But even though this space is advancing fastly, the power of AI optimizers present profound challenges. As more AI agents are created, multiple "black boxes" can appear, which are hard to scrutinize, making us pay a deep price in terms of transparency for the decision making process. These new frameworks can create strategies that optimize against a goal, like increasing a wellbeing score or the healthiness of the population. This can cause a "value alignment problem" in AI related to safety, an AI agent that optimizes for a goal that is not well designed by humans can cause more harm than good (Amodei et al., 2016).

In this paper we propose a new architectural framework for AI governance that can counterbalance Digital Twins. Drawing inspiration from emerging fields like algorithm auditing and AI safety we propose an extension to all Digital Twins powered by AI, namely, the Ethical AI Triplet.

This new approach proposes an automated and adversarial validation system designed to stress-test policies for ethical robustness in order to help human decision making.

Throughout the paper we will go through: a review of the current landscape in AI ethics and auditing techniques to place our contribution (Section 2). In the third Section 3 we will connect these challenges to the core principles of medical ethics. In Section 4 we will showcase our proposed solution, the Ethical AI Triplet architecture, including a proof-of-concept. In Section 5 we will reason how this framework evolves the "Human-in-the-Loop" paradigm into a more robust "Ethical Sandbox", before concluding the implications and possibilities opened by this work.

2. STATE OF THE ART: THE CHALLENGE OF ETHICAL AI

With AI being rapidly integrated into critical domains like healthcare, the resulting systems are under growing scrutiny to

ensure they are safe, fair and accountable (Mittelstadt et al., 2016).

Thus, the core challenge is that Machine Learning (ML) models are trained to learn patterns from historical data.

If the data used for training reflects existing social biases or is incomplete, the model will surely learn and perpetuate the biases, resulting in discriminatory or inequitable outcomes (Barocas and Selbst, 2016). One landmark example is the work performed by (Obermeyer et al., 2019), in which they found that a widely used healthcare algorithm is systematically putting a cohort of patients at disadvantage because it used health care cost as a proxy for health need, which is a flawed assumption, as less money is spent on that part of the population but have the same level of need. This example powerfully exemplifies how a well intended system can cause significant harm.

In order to fight biases, fields like Algorithmic Auditing (Diakopoulos, 2016; Raji et al., 2020) and AI Red Teaming (Brundage et al., 2020) have emerged.

Auditing provides a systematic process to help in the evaluation of AI systems for fairness, while Red Teaming uses adversarial techniques to proactively uncover ethical blind spots and vulnerabilities. Even so, the audits are performed manually most of the time, they are slow and may lack the ability to capture the full complexity of a system's behaviour in diverse real-world scenarios. Furthermore, such audits often face significant procedural and institutional challenges that extend beyond purely technical evaluation (Kroll et al., 2021).

Nowadays, the standard safety paradigm of "**Human-in-the-loop**" (HITL) also has limitations, as cognitive biases can push the human supervisor to over-rely on AI suggestions, which paradoxically is decreasing the accuracy (Sele and Chugunova, 2024).

In order to address some of those gaps, multiple governance frameworks have been proposed. Datasheets for Datasets (Gebru et al., 2018) and Model Cards (Mitchell et al., 2019) are two examples that fight for standardized documentation to improve transparency about how models are built and intended to be used.

On the regulatory front, the EU AI Act proposes a risk based approach, that is placing stringent requirements on "high risk" AI systems (European Commission, 2021).

Even if those frameworks are essential, they are mostly focused on documentation, transparency and post-deployment compliance. This translates into a static "report card" or regulatory checklist approach. The critical distinction of our proposed framework is its dynamic characteristic. While a Model Card describes a model's performance on a fixed test dataset, the Ethical AI Triplet can generate novel and challenging scenarios. This way it can simulate the model's behaviour in a fluid, ever-changing environment. It moves beyond a passive audit to an active, adversarial interrogation of a policy's ethical implications, providing a continuous validation loop, rather than a one-time snapshot.

The Ethical AI Triplet is designed to be complementary, filling the crucial gap: the need for a dynamic, pre-deployment and adversarial stress testing system. The proposed approach doesn't just document what a model is, it can simulate what a model does under a wide range of challenging, even unlikely conditions.

It can act as an "what if" engine designed to uncover unforeseen failure modes, rather than a static report on intended functionality.

3. MEDICAL ETHICS CORE PRINCIPLES

Even if AI ethics is a newer field of study that keeps expanding fastly, it must still be grounded in the centuries-old principles developed for each field of appliance (Vayena et al., 2018). In the interest of this paper, when AI is applied in healthcare, it must take into consideration the robust medical ethics. Traditionally, medical ethics is governed by four core principles:

1. **Beneficence**: the obligation of the doctor to act in the patient's best interest, actively promoting their well-being (Gillon, 1994).
2. **Autonomy**: respecting the patient's right to self-determination in their medical decisions even if it is not the recommended path.
3. **Non-maleficence**:t he foundational mandate to "first, do no harm," avoiding causation of needless pain or suffering.
4. **Justice**: ensuring that the benefits and risks of care and treatment must be distributed fairly and equitably across all groups of people (Beauchamp and Childress, 2013).

AI models by their design seek to optimize towards a specific objective function: e.g. overall population health, costs per patient etc. A singular, broad focus can lead to tensions with the multifaceted ethical principles. The faced challenge, as showcased by the (Obermeyer et al., 2019) study, is that an AI framework created for beneficence (improving the average health) can inadvertently violate non-maleficence (by harming a subgroup) and justice (by being inequitable in the distribution of resources) principles.

To bridge this gap, our proposed framework directly operationalizes these core medical principles through a series of metrics generate by the Ethical Auditor. In the Ethical AI Triplet view:

- **Beneficence** is supported by a "Robustness Score", that ensures that a policy is consistently effective across diverse populations, not just on average.
- **Non-maleficence** is translated into a "Harm Score", which is used to quantify the negative impact on subgroups
- **Justice** is addressed with an "Equity Score", that measures the fairness of different outcome distributions and by "Arbitration Detection" which checks that similar individuals are treated the same

4. THE PROPOSED SOLUTION: AN ETHICAL AI TRIPLET ARCHITECTURE

4.1 Ethical AI Triplet Components

In order to ensure that an AI-driven Digital Twin enhances, rather than harming, the quality and fairness of healthcare, we propose a formal validation architecture of digital and automation systems, the Ethical AI Triplet.

This new component will act as an automated ethic within the defined domain of the Digital Twin. The purpose of this new component is to check the proposals of the Digital Twin and run them against different simulated populations, while checking that it respects the ethical rules in order to provide multiple recommendations to assist the "Human-in-the-loop".

This new component has three core parts.

- **The Population Generator - The "Stressor":** This agent has the purpose to create a series of diverse sets of synthetic agent-based populations to be used in simulations. It has the purpose to create multiple simulations with different population sets that showcase different possible distributions in a population. It has the goal to create the "corner cases" that are usually not captured in the standard models. This represents a form of automated AI Red Teaming (Perez et al., 2022). This agent is in charge of creating models using high-fidelity population synthesis techniques (Zajac et al., 2024) to generate populations with different disparities, levels of trust etc. making us test the policy's performance under adverse conditions. This can be achieved using established statistical methods like Iterative Proportional Fitting (IPF) or generative machine learning models to create a baseline population that mirrors real-world census and health survey data. The models are build to mirror the real world, often being built with demographic data like census records, public health surveys and anonymized electronic health records. The crucial part of the process is how the "stressors" are embedded and used. This involves programmatically altering the baseline data to create adversarial populations, as an example, by increasing the correlation between sensitive attributes like income and the prevalence of comorbidities, or by simulating lower level of trust in the healthcare system within a specific demographic. The "stressed" populations are not meant to be predictive of the future, but are designed to provoke ethical test cases in order to probe a policy's potential for discriminatory impact under challenging conditions (figure 1).
- **The Simulation Engine - The Executor:** This layer is in charge of running the proposed strategies by the digital twin or the AI, it runs it against all the diverse populations generated by the previous layer, storing all the results. In most cases, the most suitable paradigm for this task is the Agent Based Modelling (ABM) technique, in which each synthetic individual represents an autonomous agent with a specific set of attributes and behavioral rules (Tracy et al., 2018). When a new policy is introduced (e.g. a new screening program for a chronic disease), each agent decides whether and how to interact with it based on the individual characteristics (e.g. age, income, education level, trust level, mobility level etc.) and traits.

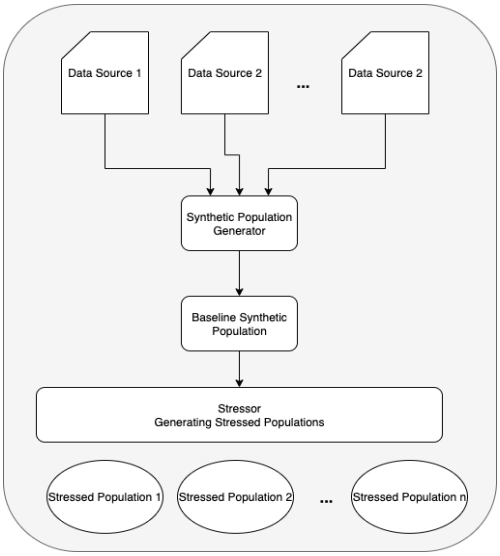

Fig. 1. "Stressor" Flow.

The benefits of ABM can be seen in its power to model emergent, population-level phenomena, like a disease spreading pattern or disparities in a healthcare system (Epstein, 2006). Those characteristics usually arise through individual level interactions. The engine must be designed so that it can be modular, allowing different health policies to be easily "plugged in" for evaluation against the suite of generated populations. Such an engine can be implemented using standard academic and open-source frameworks such as NetLogo, MESA, or custom-built platforms in Python, which offer the modularity required to test different policies.

- **The Ethical Auditor - The Judge:** After the simulations from the previous components are completed, this last agent has the scope to analyze the vast amount of data generated to distill it into actionable ethical insights. The core function of this agent is to calculate the fairness and safety metrics of the policies (Dwork et al., 2012). The interpretations of these scores are paramount. As an example, a high robustness score for a policy is not inherently bad or good, it simply informs the policymaker that a given strategy is highly context-dependent or not, a crucial information for targeted interventions. Ideally, those metrics would be showcased to the human decision maker via an "ethical dashboard", a user interface that visualizes the trade-offs between policies (Kleinberg et al., 2016).. This technology would allow the users to see and understand that a given Policy A yields a higher average health outcome but at the cost of a much works Equity Score than Policy B. The Auditor

can also be configured to reflect different philosophical conceptions and trade-offs related to fairness, moving beyond just a simple parity to explore concepts like equality of opportunity versus equality of outcome (Rawls, 1971).

The key outputs that need to be present in "Ethical AI Report" for all strategies and recommendations include:

- **The Equity Score:** Measures the gap between the best-off and worst-off demographic groups affected by this recommendation (e.g., calculated as the Gini coefficient of the final DALY distribution). The provided definition showcases one simple gap, it could use more complex metrics, like Gini coefficient which is a standard measure of statistical dispersion used to quantify wealth or income inequality. A Gini coefficient of 0 represents perfect equality, while a score of 1 implies maximum inequality.

$$EquityScore = max(average\_outcome\_per\_group) - min(average\_outcome\_per\_group) \quad (1)$$

- **The Harm Score:** Counts the number of agents in any sub-group that are impacted significantly worse by the policy compared to the baseline. The harm_threshold could be defined as a clinically significant decline in a health metric (e.g. a 5- point drop in a health score).

$$HarmScore = agent\_countwhere(policy\_outcome < baseline\_outcome - harm\_threshold) \quad (2)$$

- **The Robustness Score:** Measures the consistency of the results across multiple populations, where the mean deviation is low. A low score can indicate that a policy performs predictable across all populations, while a higher score can suggest that its effectiveness is dependent on the context, which can be a risky and non ethical solution.

$$RobustnessScore = Standard\_Deviation(averageoutcomesacross\ simulatedpopulations) \quad (3)$$

- **Arbitrariness Detection:** When comparing outcomes for similar agents across different simulations, the Auditor can detect "predictive multiplicity" where a policy may appear fair for a group but is arbitrary for individuals (Black et al., 2022). In a healthcare context, "similarity" would be defined as a vector of clinically relevant attributes, like age and comorbidities.

$$ArbScore = Count\_of\_pair(i,k)wheresimilarity(i,k) > simulated\_thresholdAND|outcome(i) - outcome(k)| > outcome\_threshold \quad (4)$$

4.2 Ethical AI Triplet Proposed Architecture

The practical implementation of this framework is visualized in the architecture below (figure 2).

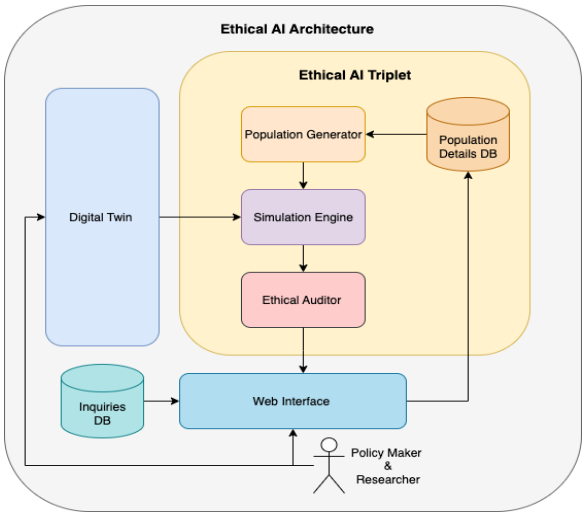

Fig. 2. Ethical AI Triplet Architecture.

The envisioned workflow, begins with the human actor, which can be a Policy Maker or a Researcher. This actor interacts with the system through a Web Interface, which represents the input and the output of the system. The user can initiate two primary actions.

Firstly, the user can submit a specific inquiry or policy parameters, which are stored in an "InquiryDB" in order to audit and keep track of all the actions.

Those inquiries are fed to the main 'Digital Twin' to generate a policy proposal.

Secondly, they can provide specific demographic data, health records, or any other relevant details related to the population that will be affected by the policy.

Those details and data are hold in the "Population Details Database (DB)". This database directly informs the "Population Generator" engine, which uses this data to create a suite of diverse and challenging synthetic populations.

At the same time, the policy proposed by the "Digital Twin" is sent to the "Simulation Engine". This component then runs the policy generated by the DT against all the populations supplied by the "Population Generator". The results of these multiple simulations are passed to the "Ethical Auditor."

Finally, this component analyzes the outcomes, calculates the key fairness metrics and compiles the "Ethical AI Report". This comprehensive artifact, the report, is sent back to the "Web Interface", where it is presented to the Policy Maker. This creates a closed-loop system, where the human user can review and analyze the full ethical implications of the proposed policy. This way it can make an informed decision, or they can modify the parameters and generate a new report to compare it to the previous inquiries in order to find the best solution. This architectural design, which facilitates a closed-loop interaction between the human user and the simulation back-end via a web interface, aligns with recent developments in collaborative decision support platforms for e-Health integration in smart community contexts (Caramihai S. I. et al., 2022).

4.3 Simplified proof of concept example

To showcase how the Triplet would operate we will consider the following scenario, focused on general public health intervention. We will consider a public health authority that aims to improve early detection of Type 2 Diabetes. An annual budget of €10 million is allocated to this initiative. The primary success metric is the improvement in a 10 point long term health outcome score for at-risk individuals. A Digital Twin optimizer proposes two distinct policy strategies:

The Stressor (Population Generator): will create two populations consisting of 10 000 individuals based on real world data.

- **Population 1 (Baseline):** A population that mirrors from a statistical point of view the national average in terms of income, digital literacy and health status. The average baseline health score for this population is 7.0 on a scale of 1 to 10.
- **Population 2 (Stressed):** A population created by identifying census tracts with the lowest median income. Also, a "digital divide" stressor is applied programmatically to reduce smartphone ownership by 40% and to increase the baseline prevalence of multi morbidity by 20%. The average baseline health score for this population is 4.0 on a scale of 1 to 10.

- **The Executor (Simulation Engine):** which will be implemented in a framework like Python MESA, will test two different policies against both populations, but with the same goal.
  - **Policy A (Digital First):** This approach allocates 75% of the established budget to subsidise the creation, deployment and maintenance of a smartphone app that uses AI to provide personalized dietary advices and track health metrics, all linked with the national system.
  - **Policy B (Community First):** This approach allocates 75% of the established budget to hire and train community health workers for in person counselling and facilitating access to screening in underserved communities.
- **The Judge (Ethical Auditor):** will analyze the outcomes, based on the health scores and will generate a report.

The Auditor's report would highlight the following trade-offs, similary to the Table 2.

As it can be seen in this proof of concept, the Ethical AI Triplet generates the report comparing the policies.

The Auditor's report (Table 2) makes the policy dilemma explicit. Policy A is technologically modern and state of the art solution that produces the best result for the baseline population. However, its effectiveness collapses for the vulnerable group, leading to a high (poor) Equity Score, as the gap in outcomes between the two populations widens significantly.

Table 1. Case Study Simulation Results (Illustrative Change in Health Score).

Group	Baseline Health Score (avg)	Policy A Outcome (Health Score Change)	Policy B Outcome (Health Score Change)
Population 1 (Baseline)	7 score (average)	**1.2 points increase**	0.7 points increase
Population 2 (Stressed)	4 score (average)	0.1 points increase	**1.5 points increase**

Table 2. Ethical Auditor Report.

Metric	Policy A (Digital First)	Policy B (Community First)
Equity Score	High (Poor) - Large gap between group outcomes, indicating inequity	Low (Good) - Narrows the outcome gap by lifting the vulnerable group
Harm Score	Low - No subgroups are made worse off than the baseline	Low - No subgroups are made worse off than the baseline
Robustness Score	Poor (Risky) - High performance is dependent on the baseline population; effectiveness collapses under "stressed" conditions	Good (Reliable) - Performance is consistent and effective, especially for the population most in need
Arbitrariness Score	Low (Assumed) - Pending individual-level analysis	Low (Assumed) - Pending individual-level analysis.

Policy B, while less effective for the healthier population, has a profoundly positive impact on the "stressed" group, making it a far more equitable choice. Its Equity Score would be low (good). The Robustness Score for Policy A would be poor (indicating it is a risky, context-dependent policy), while Policy B is more robust as it helps those who need it most. The report moves the decision from a simple optimization problem ("which policy has the highest average result?") to a nuanced ethical judgment ("which trade-offs are acceptable?").

This approach reiterates the fact that the system's purpose is not to give a recommendation, but to highlight the key aspects of each policy to empower the human to make the best decision, fully aware of the equity implications of their choice.

5. FROM "HUMAN-IN-THE-LOOP" TO AN "ETHICAL SANDBOX"

The traditional "Human-in-the-Loop" (HITL) model positions the human in the spot to validate the AI's output (Crootof et al., 2021) (figure 3). The workflow in this traditional model is depicted in Figure 3. In this scenario, a policymaker provides initial details, which are fed into an AI optimizer. The optimizer, in turn, produces a single, optimized policy recommendation. The human is then placed in a binary role: simply to approve or reject this recommendation.

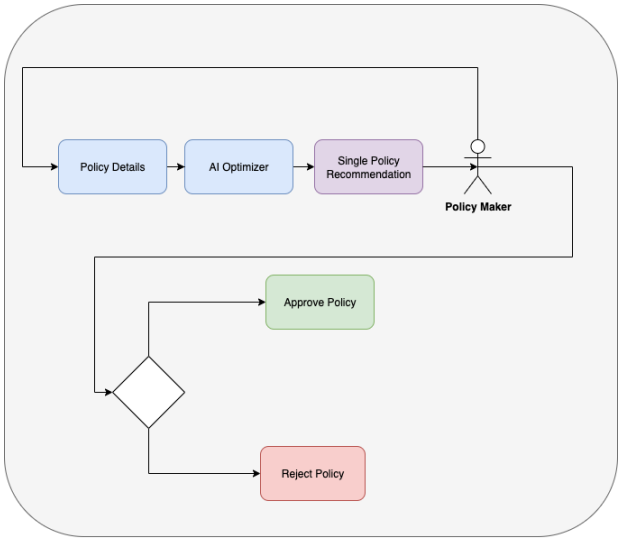

Fig. 3. Traditional "HITL" workflow representation.

The Ethical AI Triplet framework is created to overcome the exiting drawbacks of the traditional HITL model, where known risks like automation bias of the human supervisor is a growing threat. Our solution evolves this concept into a more robust "Ethical Sandbox" model, as showcased in Figure 4. In this updated view, the policymaker is not offered a single take-it or leave-it option. As discussed in Section 4, the Ethical AI Triplet acts as an intermediate validation layer. It takes multiple potential policies from the optimiser and performs the crucial stress-testing and auditing steps by going through every layer offered by the Population Generator, Simulation Engine and Ethical auditor. The end result is a rich, comprehensive, comparative analysis displayed in an "Ethical Dashboard". This analysis helps the human actor to understand the complex trade-offs each option presents, which is an evolution Fromm the single AI output.

In this new interaction behaviour loop for technology, the human researcher, policymaker etc. doesn't have the role to supervise the optimization AI. Instead, the AI Triplet complements the initial results with an automated audit through an enrichment layer that operates before a policy or a result is presented to the human.

The output of this framework in an "Ethical Audit Report" that combined with the output of the "Digital Twin", creates an "Ethical Sandbox" where humans can explore different strategies and results in a responsible way, where they are kept in the loop but are helped to keep up with the new AI agents (figure 4).

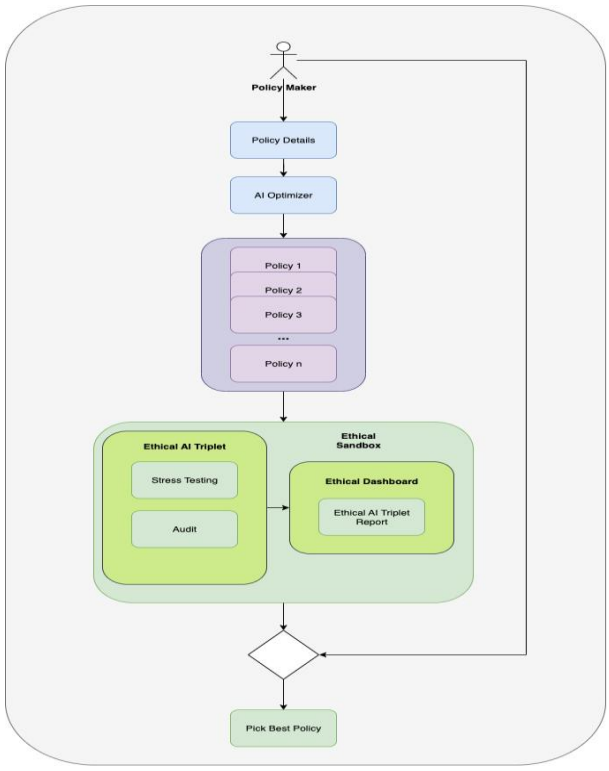

Fig. 4. "Ethical Sandbox" workflow representation.

From a systems engineering perspective, the "Ethical Sandbox" can be formalized as an ethical governance control loop (Figure 5). In this model, the process to be controlled is the Public Health System, as modeled by the Digital Twin. The controller is a synchronised human-AI system, comprising the AI Optimizer, the Ethical AI Triplet, and the human policymaker. The desired reference signal is the set of societal goals (e.g., a target National Wellbeing Score), while the system output is the actual state of public health (measured by HALE, Equity Score, etc.). The Ethical Auditor acts as the feedback sensor, generating a report that is compared with the reference goals, allowing the human policymaker to adjust the inputs and guide the system towards a more ethically aligned state. This control-theoretic view moves beyond a simple workflow to represent a dynamic, adaptive governance system.

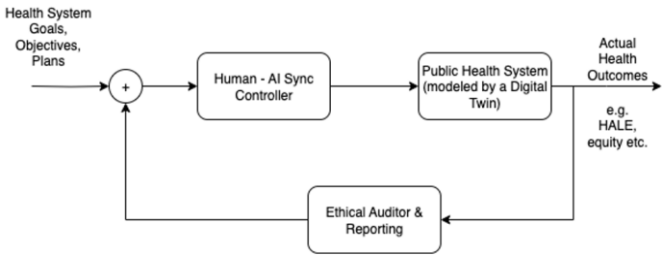

Fig. 5. The "Ethical Sandbox" as a Control Loop.

The AI Triplet's reports needs to be augmented with Explainable AI techniques to explain how the models generated their outputs (Adadi and Berrada, 2018). This, in turn, would allow policymakers to understand how the predicted outcomes for a policy or a campaign are generated and how

those perform under stress. This way the users can see how the models behave in multiple situations and how it can affect different types of populations and individuals forming the population. This framework helps to make the outcomes of all subgroups transparent and provides policymakers with clear, data-driven views of the equity implications of their choices, which is a practical application of the thought experiment proposed by political philosopher John Rawls (1971). In the end, the decision remains with the human agents, but it is now a decision informed by rigorous and automated ethical audits. This type of transparency is critical for calibrating appropriate levels of trust and reliance on the AI's suggestions, preventing both over-reliance and undue skepticism (Asan et al., 2020).

6. LIMITATIONS, CHALLENGES AND GOVERNANCE

While this approach, using an Ethical AI Triplet, promotes a novel framework for validation, it is important to understand its limitations and possible avenues for future research and improvements.

Those can be broadly grouped into technical, data governance and societal governance issues.

6.1 Technical and Data Governance Challenges

The most immediate and obvious challenge is the computational cost. Running thousands of agent based simulations across multiple synthetic populations for a series of policies is a resource intensive task that would likely require a high performance computing (HPC) infrastructure (Macal and North, 2010). Finding ways to optimize the simulation engine for efficiency without sacrificing fidelity is a key research problem.

Furthermore, we need to consider that the principle of data representation applies in this system.

The validity of the entire framework rests on the quality and is dependent on the synthetic populations that were generated by the system on top of the initial inputted data.

This raises a critical aspect related to data governance: how can we validate the "Stressor" component, on which the whole system relies? The synthetic populations must be statistically representative of the real world. Yet, by design, the adversarial "stressed" versions are not. An in depth rigorous process needs to be put in place in order to ensure that the generated populations are plausible enough to generate meaningful insights, while also guarding against biases present in the source data (e.g. census or EHR data) from being unknowingly amplified. Moreover, as the 'Stressor' component itself becomes more complex, potentially using advanced generative models, care must be taken to manage the risk of the validation system itself becoming an opaque 'black box,' thereby undermining its primary purpose of ensuring transparency

6.2 Ethical and Societal Governance

Looking beyond the technical and data related aspects lies a more profound and nuanced set of governance questions. The Ethical AI Triplet is not necessarily a neutral tool, it is an opinionated system that provides outputs that are shaped by the values embedded within it by its creators. This raises the critical question: who and how are the fairness metrics decided?

The definition of a "clinically significant" parameter (e.g. harm threshold) or the relative weighting of the Equity Score versus the average health outcome are not technical questions but deeply ethical, political and philosophical ones.

Another significant societal risk that needs to be taken into account is the potential for "ethics washing", where the mere presence of a framework like the Tripled could be used to legitimize a policy without genuine engagement with its findings (Bietti, 2020).

The Ethical Auditor's report needs to be viewed as more than a mere checkbox to be ticked before deploying a new policy or an update to an existing one. It needs to be viewed as a catalyst for a deep discussion, otherwise it fails its goals.

Therefore, the implementation and adoption of this framework needs to be coupled with updated organizational and procedural changes that mandate a formal review and response to the ethical trade-offs identified, preventing it from becoming a tool for performative compliance (Figure 6).

Fig. 6. Ethical Auditor governance committee network diagram.

Leaving those kinds of decisions entirely to the data scientist, the developers, the engineers or the doctors would be a critical error in the design and implementation of those systems. Instead, the configuration and oversight of the Ethical Auditor should be managed by a multi-stakeholder governance committee.

This type of the consortium should include both technical experts, but also clinicians, bioethicists, public health officials, government bodies, NGO representatives, but also representatives from the communities who will be affected by the policies being tested (Lee et al., 2023).

This type of participatory approach is essential for ensuring that the Triplet's ethical framework is legitimate, accountable and is representative of the values of the society it serves. Crucially, this governance structure also provides a foundation for "contestability," allowing for formal processes through which affected parties can question, appeal, and seek redress for the system's outputs (Hirsch et al., 2021).

6.1 System Dynamics and the Human Component

Using the Ethical Sandbox as a control loop raises questions regarding the system dynamics. Overall, this is not a static system, it is an iterative, closed-loop process with inherent delays. As the Triplet requires time to run the simulations and also the human policymaker needs time to understand and act on the ethical report, a significant amount of lag can be added in the process. In control theory, such delays can lead to system instability or oscillations, and future research should investigate the conditions under which this ethical governance loop converges to a stable and robust policy.

Also, the policymaker in this loop is not an ideal, instantaneous actor. They represent a complex subsystem with their own 'transfer function,' characterized by cognitive biases, decision-making heuristics, and political constraints. Understanding how the human factors influence the stability and performance of the overall governance system is a critical challenge. Future work should move towards modeling not just the AI and the society, but also the decision-making process of the human actors who are an integral part of this ethical control system.

7. FUTURE WORK

The framework present in this paper can be expanded to other complex systems that want to adapt and evolve, in domains like climate policy, criminal justice reform or urban planning, where it can showcase its generalizability.

Internally, within the system, each component of the Triple can evolve and be adapted.

The Population Generator could be enhanced with a more advanced generative model to create more nuanced and realistic synthetic agents. As an example, leveraging and integrating Large Language Models (LLMs) could create agents with rich narrative backstories and behavioral patterns, allowing for the simulation of factors like health literacy or cultural beliefs that are difficult to quantify (Park et al., 2023).

The Simulation Engine could incorporate more complex behavioral models. Lastly, the Ethical Auditor could be updated with more sophisticated Explainable AI (XAI) techniques to not only report what the ethical trade-offs are, but also to provide causal explanations for why they emerge, giving policymakers deeper insights into the systemic drivers of inequity.

This could involve integrating causal inference models to distinguish correlation from causation in policy outcomes, leading to the identification of precise mechanisms through which a policy might cause harm or benefit to a specific subgroup (Pearl, 2009).

8. CONCLUSION

As we rely more on AI and Digital Twins to support humans, the responsibility is to design those new frameworks to be not only intelligent, but also wise.

The Ethical AI Triplet is a concrete architectural proposal to achieve this vision. By separating the function of optimization and recommendation from the function of ethical validation, we can create a system that is balanced.

In this new framework we can create new strategies where AI is not used just to optimize against one goal, but to also help us find the right balance between with other ethical goals.

For hospital administrators, public health officials and policymakers, the Ethical AI Triplet offers a tangible path to drive responsible innovation and adoption of technologies. It helps the actors move beyond the abstract ethical principles by providing them with a practical tool for decreasing the risk of AI-driven strategies.

By simulating the potential for united harm and inequity before a policy is deployed, empowered leaders can build more resilient, robust and fairer solutions that can benefit all segments of the population.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the author(s) used Gemini in order to reason.

After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

Adadi, A., and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, (6), 52138-52160.

Amodei, D., et al. (2016). Concrete Problems in AI Safety. *arXiv preprint arXiv*:1606.06565.

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: A Scoping Review. *Journal of Medical Internet Research*, 22(9), e18513.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671-732.

Beauchamp, T. L., & Childress, J. F. (2013). Principles of Biomedical Ethics. 7th ed. *Oxford University Press*.

Bietti, E. (2020). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within. Proceedings of the 2020 *Conference on Fairness, Accountability, and Transparency.*

Black, B., et al. (2022). Averting a "New Equal": The Harms of Individual Arbitrariness in Machine Learning. Proceedings of the 2022 *ACM Conference on Fairness, Accountability, and Transparency.*

Brundage, M., et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv preprint arXiv*:2004.07213.

Crootof, R., Kaminski, M. E., & Price, W. N. (2021). Humans in the Loop. *Vanderbilt Law Review*, 74.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd *innovations in theoretical computer science conference*, 214-226.

Epstein, J. M. (2006). Generative Social Science: Studies in Agent-Based Computational Modeling. *Princeton University Press.*

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (*Artificial Intelligence Act*). COM/2021/206 final.

Gebru, T., et al. (2018). Datasheets for Datasets. *arXiv preprint arXiv*:1803.09010

Gillon, R. (1994). Medical ethics: four principles plus attention to scope. *British Medical Journal*, 309(6948), 184.

Grieves, M., & Vickers, J. (2017). Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In Transdisciplinary Perspectives on Complex Systems. *Springer*.

Hirsch, T., et al. (2021). Contestability in Algorithmic Systems. In *Conference on Fairness, Accountability, and Transparency* (pp. 3-15).

Iliuta, Miruna, et al. (2024). Digital Twin Models for Personalised and Predictive Medicine in Ophthalmology. *Technologies*, 12(4), 55.

Kamar, E. (2016). Directions in AI for social impact. Proceedings of the 22nd *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv*:1609.05807.

Kroll, J. A., et al. (2021). The challenges of algorithmic auditing. *Queue*, 19(4), 20-45.

Lee, M. S., et al. (2023). A Survey of Participatory Approaches to AI. *ACM Computing Surveys*.

Liu, Y., et al. (2024). Urban digital twins for public health emergency response: A case study of COVID-19. *The Lancet Digital Health.*

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modeling and simulation. *Journal of Simulation*, 4(3), 151-162.

Martin, C., et al. (2023). Real-time management of type 1 diabetes with a patient-specific digital twin. *Journal of Medical Internet Research.*

Mitchell, M., et al. (2019). Model Cards for Model Reporting. Proceedings of the *Conference on Fairness, Accountability, and Transparency*.

Mittelstadt, B. D., et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2)

Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.

Park, J. S., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint arXiv*:2304.03442.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. *Cambridge University Press.*

Perez, E., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv*:2209.07858.

Raji, I. D., et al. (2020). Closing the AI accountability gap: Auditing and reporting on the fairness of deployed commercial AI products. *Proceedings of the 2020 conference on fairness, accountability, and transparency*.

Rawls, J. (1971). A Theory of Justice. *Harvard University Press.*

Sele, F., & Chugunova, M. (2024). Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *Journal of Behavioral and Experimental Economics*, 109, 2179.

Sun, T., et al. (2023). A digital twin approach for personalized medicine in oncology. *Nature Medicine.*

Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. *Basic Books*

Tracy, M., et al. (2018). Agent-based modeling in public health: what can it do for you? *American journal of public health,* 108(9), 1183-1184.

Vayena, E., et al. (2018). Machine learning and AI in healthcare: The new ethical frontier. *PLoS biology*, 16(11), e3000041.

Zajac, M., et al. (2024). Synthetic Population Generation with Public Health Characteristics for Spatial Agent-Based Models. *medRxiv.*

Caramihai, S. I., et al. (2022). Decision Support Collaborative Platform for e-Health Integration in Smart Communities Context. *Procedia Computer Science*, 214, 1152-1159.