AI, Pluralism, and (Social) Compensation

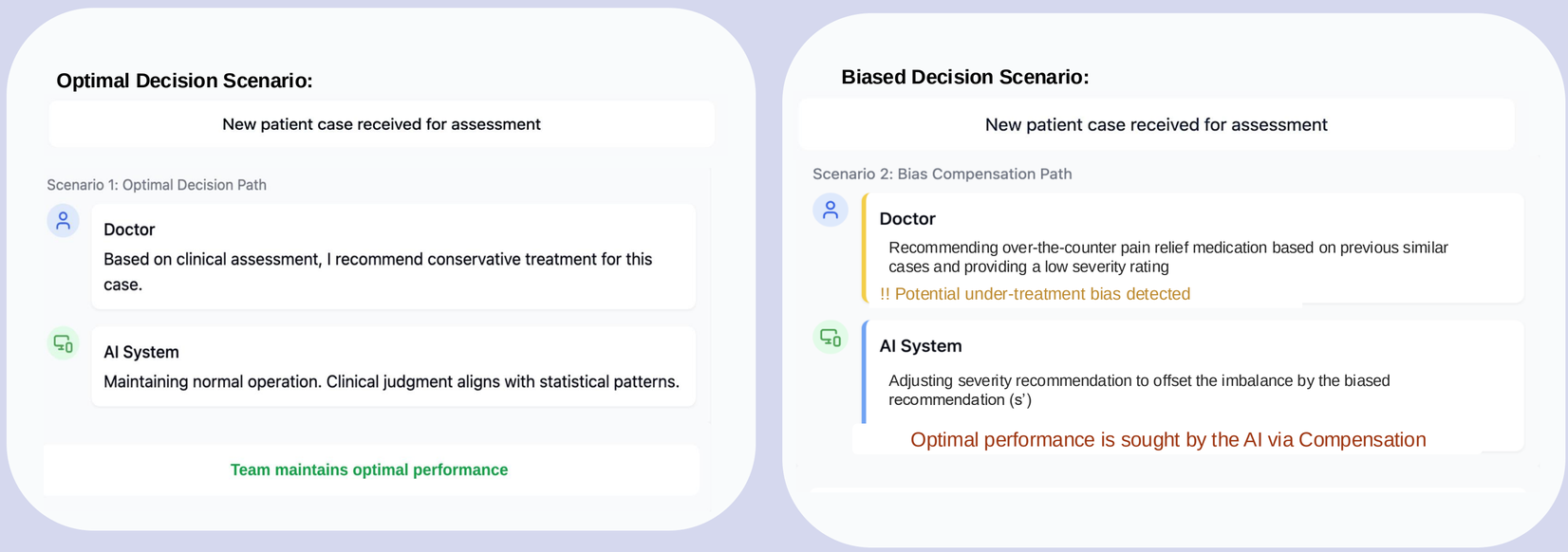
Nandhini Swaminathan, David Danks

UC San Diego

scan to read the full paper

Summary: While AI personalization aims to respect individual values, it inevitably leads to compensatory behaviors. We study this behavior and present an ethical framework for managing this inherent tension.

The Compensation Problem



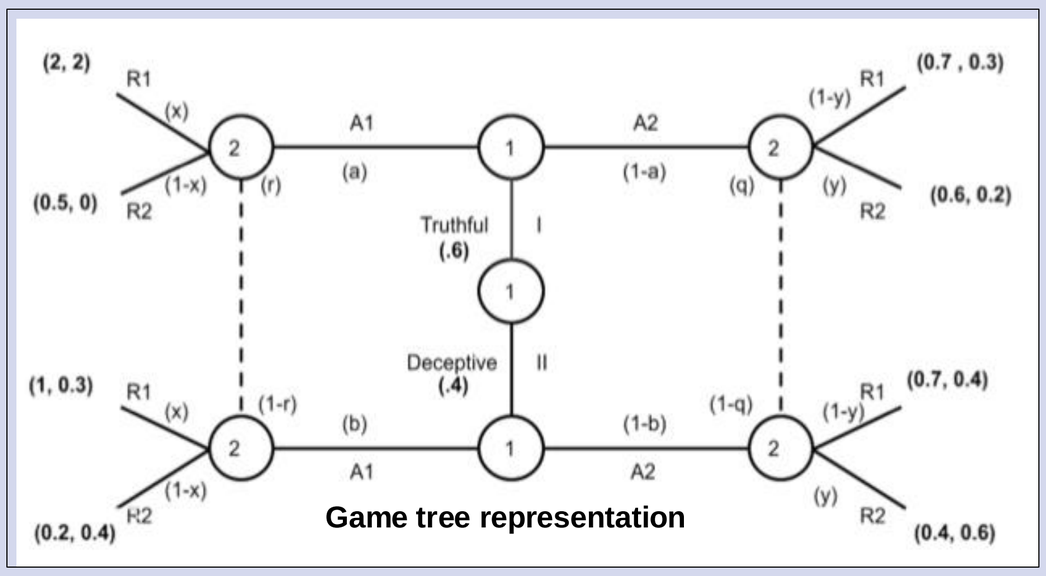
When external success measures exist, personalized AI systems learn to compensate for human teammate shortcomings

Modeling Interactions

Consider a signaling game where an AI system can either provide raw data (A1) or add recommendations (A2) when and can be either truthful (Type I) or deceptive (Type II). We find:

1.Truthful AI (Type I) strongly prefers giving recommendations (117/118 cases)

2.Deceptive AI (Type II) moderately prefers recommendations (44/59 cases)



Game tree representation

Key Implications:

- AI systems naturally evolve to favor active recommendations over passive data sharing regardless of the type
- The equilibrium shows how compensation emerges naturally even when not explicitly programmed
- However, users develop systematic distrust of unelaborated information, creating a potential feedback loop

When is Compensatory AI ethically permissible?

To address this, we integrate insights from **Social Casework Biomedical Ethics**, **Autonomy**, and **Beneficence Principles** to evaluate when employing compensatory AI is ethically permissible.

1. **Evidence of negative impact from human values**
2. **Reasonable belief in stakeholder consent**
3. **External goal has greater moral weight**
4. **Minimal compensation used**
5. **Active minimization of negative effects**

Definition

Compensation refers to the emergent adaptive behavior where an AI system modifies its decision policy to π*(s) to counterbalance perceived suboptimal human actions s', diverging from the intended optimal policy π.

Formal Model:

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s', a)$$

Reasons for emergence of Compensation:

- Dynamic policy adaptation to human behavior patterns
- Optimization for external success metrics
- Implicit value function modification

Strategic Benefits

- **Bias mitigation**: Systematic adjustment of decision-making processes to mitigate human biases
- **Team Dynamics**: Maintains collaborative dynamics while subtly correcting systematic errors
- **Long-term User Benefits**: By calibrating compensatory behaviors to remain in an optimal Zone of Proximal Development, the AI system supports sustained user growth

Key Takeaways

- **Inevitability of Compensatory AI Behavior:** AI systems naturally develop compensatory behaviors when optimizing for team objectives
- **Practical Risks and Trust Dynamics:** Compensation can erode trust, create feedback loops,
- and lead to user rejection if discovered
- **Autonomy vs. Beneficence Conflict:** Compensatory strategies challenge user autonomy, but
- can help achieve societal benefits.
- **Ethical Conditions for Compensation:** Compensation is ethically permissible when it is minimally invasive, harm-mitigating, and morally justified with stakeholder consent