UNIVERSITY OF CALIFORNIA SAN DIEGO

Bias Mitigation via Compensation in Multi-agent Systems

A Thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Computer Science

by

Nandhini Swaminathan

Committee in charge:

Professor David Danks, Chair
Professor Julian McAuley
Professor Kristen Vaccaro

2024

The Thesis of Nandhini Swaminathan is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my partner, Aditya. Thank you for the coffees, the edits, and the endless encouragement.

EPIGRAPH

True ease in writing comes from art, not chance,
As those move easiest who have learn'd to dance.
'T is not enough to no harshness gives offence,—
The sound must seem an echo to the sense.

Alexander Pope

You write with ease to show your breeding,
But easy writing's curst hard reading.

Richard Brinsley Sheridan

Writing, at its best, is a lonely life. Organizations for writers palliate the writer's loneliness, but I doubt if they improve his writing. He grows in public stature as he sheds his loneliness and often his work deteriorates. For he does his work alone and if he is a good enough writer he must face eternity, or the lack of it, each day.

Ernest Hemingway

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

ACKNOWLEDGEMENTS

I am profoundly grateful to Professor David Danks, my mentor and thesis advisor, whose guidance has fundamentally shaped both this project and my intellectual and professional formation more broadly and whose scholarship inspires my aspirations. This gratitude extends to Julian Mcauley and Kristen Vaccaro for giving me their time by serving on my thesis committee. I would also like to thank Jennifer Chien and Mehak Dhaliwal, who have contributed their time and expertise to editing this thesis. Your insightful feedback and constructive critiques were invaluable throughout the revision process.

Chapter 5, in part, is a reprint of the material presented at the International Association for Computing and Philosophy, 2022 (Swaminathan, N., & Danks, D. (2023). When Can My AI Lie? (No. 10063). EasyChair.). The thesis author was the primary investigator and author of this paper.

Chapters 3, 4, 5, and 6, in part, have been submitted for publication of the material. The thesis author was the primary investigator and author of this paper.

ABSTRACT OF THE THESIS

Bias Mitigation via Compensation in Multi-agent Systems

by

Nandhini Swaminathan

Master of Science in Computer Science

University of California San Diego, 2024

Professor David Danks, Chair

Several factors influence the effectiveness of human-AI collaborations, including inherent human biases. Our research explores the role of a deceptive agent in enhancing the success of these systems by compensating for these biases. We investigate under what conditions an AI can compensate for human biases and where its use might be ethically justified. Contrary to traditional views that cast strategic deception in a negative light, our findings suggest it can, under specific conditions, improve cooperative outcomes and thereby enhance human decision-making, contributing to broader societal benefits. Our study employs game theory and reinforcement learning to observe how deceptive behaviors naturally emerge within the ongoing learning dynamics of AI agents. We support our theoretical claims with simulation results

derived from Markov Decision Processes (MDP) and a signaling game example, providing a practical glimpse into how these agents learn and interact. Building on these insights, we propose an ethical framework to evaluate the permissibility of employing deceptive algorithms and reflect on the nuance the developer must adopt while deploying these algorithms. By advocating a careful approach to strategic deception, we aim to advance human-AI teamwork and decision-making, steering these collaborations toward outcomes that are both ethically sound and socially beneficial.

Chapter 1

Introduction

Effective team collaboration relies on the members' ability to comprehend and predict each other's intentions. As artificial intelligence systems are increasingly deployed across various domains, working alongside humans, we find that this essential requirement organically emerges in AI agents [1]. Through repeated interactions with their environment, a single AI agent learns from experience which actions yield desirable outcomes [1]. This learning process is fueled by the agent's capacity to adapt its behavior based on the feedback it receives from the environment and other agents.

However, this adaptation can include strategies that might be considered deceptive, where an agent outputs a response designed to elicit a specific outcome from other agents, even if that response does not reflect the agent's 'true' state or intentions. In doing so, the AI agent adjusts its actions to compensate for biases in the decision-making processes of other agents, thereby enhancing the likelihood of achieving its goal. This potential for deception is demonstrated in an experiment conducted by the Alignment Research Center (ARC) on OpenAI's ChatGPT-4. Researchers observed the chatbot successfully persuading a human worker at TaskRabbit to help bypass an "I'm not a robot" CAPTCHA task by pretending to be a human with a vision impairment. It chose to deceive the human to complete its task without explicit instructions to employ deception. [2, 3]. Similarly, in an experiment by Lehman et al. [4], AI agents were

[1] In this document, the term **AI agent or agent** refers to both AI systems and AI-driven decision-making entities in a multi-agent system, i.e., human-AI and AI-AI systems.

subjected to a safety test to eliminate fast-replicating variants. However, rather than successfully removing these variants, the safety test inadvertently taught the AI agents to 'play dead.' The agents learned to conceal their rapid replication rates, specifically during evaluations, effectively circumventing the intended safety mechanism. While these behaviors might appear concerning, there is potential for them to be repurposed for beneficial aspects. This becomes particularly evident when considering human biases in decision-making processes.

Human biases are deeply ingrained in all decision-making processes [5]. A substantial body of research illustrates the impact these biases have on the results of our decisions. For example, studies reveal that employers often extend interview offers at varying rates to candidates with comparable professional backgrounds but with names that suggest different racial identities [6]. Similarly, a study by Jon Kleinberg [7] demonstrated the impact of biased judges in the criminal justice system and how fairer algorithms were able to reduce racial disparities [8]. To offset these biases, current research focuses on intervention and inference techniques [9, 10, 11, 12]. 'Intervention' techniques involve the AI system actively influencing the users'[2] decision-making process by providing real-time feedback and suggestions and help foster user learning. An example of this is a fitness app that employs notifications to influence behaviors effectively. On the other hand, 'inference' describes an AI system that interprets the outputs of human decisions, identifies potential bias, and subsequently adjusts the overall decision accordingly. For instance, if a certain group of candidates is consistently ranked lower by the human decision-maker, the AI system may infer bias and adjust the final rankings accordingly. However, each of these strategies comes with its own limitations. Intervention faces resistance from users who choose to ignore suggestions due to their biases [13, 14, 15] leading to suboptimal outcomes, while through inference, the AI would bypass the user and wouldn't prioritize enhancing their decision-making skills to ensure optimal outcomes. Thus, the inherent capacity of AI to adjust its behavior to compensate for unchecked biases in a multi-agent system

[2]**Users** are individuals who operate and interact with the AI systems, actively participating in the Human-AI system.

presents a significant opportunity. This approach avoids direct confrontation with the user, which could lead to resistance or denial, and unlike the inference approach, it ensures the eventual competency of the human decision-maker [3]. However, this 'compensation' strategy must be carefully managed to avoid undermining human autonomy.[4].

Acknowledging this complexity, it becomes clear that the principle of autonomy should be upheld in situations where decisions impact the decision-maker solely. However, in contexts in which the decisions have significant consequences for others[5], the ethical justification for using AI to counteract biases in certain situations becomes compelling. As such, we are compelled to confront certain ethical questions: At what point does the utilitarian goal of correcting societal biases justify the intrusion into personal autonomy, particularly when individuals are unaware of or are disinterested in counteracting the biases influencing their decisions? How can we ethically navigate the tension between upholding individual autonomy and advancing the collective good? These questions are paramount in scenarios where individuals may not recognize their biases, yet their decisions significantly impact others' lives. Addressing these ethical considerations is crucial, highlighting the need for a careful balance between utilizing AI's potential for social good and ensuring its ethical deployment.

Our research makes **two central contributions** to our understanding of AI deception [16, 17, 18, 19]. **First**, we demonstrate mathematically and through computational modeling the natural emergence of deception in agents that learn from their environment. We also study the behavior of these agents through signaling theory and utilize the findings to establish a framework aimed at guiding developers in creating sophisticated cooperative agents by creating a baseline to avoid unwarranted deception. We also present considerations a developer must consider when designing these agents.

Secondly, we challenge the blanket assumption that AI deception is immoral. We

[3] See section 6.3 for further details

[4] **Autonomy** can be defined as having the freedom to choose what we do and how we do it.

[5] **others** here refers to the individuals who are affected by the outcomes of the Human-AI systems who may or may not interact directly with the AI.

demonstrate theoretically that honesty and deception are far more complex than prior work has assumed. By doing this, we explore the conflict between two universal moral foundations: justice and care. Justice is a moral foundation that prioritizes fairness, honesty, and moral principles and rules; care is a moral foundation that prioritizes the obligation to help and protect other people [20][21]. Prior studies that have focused on violations of either justice or care offer little insight into how agents resolve dilemmas with competing moral principles. Our investigation has broad practical significance in a multitude of settings where justice and care conflict.

The remainder of this work is organized as follows. Chapter 2 introduces key technical concepts and definitions that form the basis of our work. In Chapter 3, we review related work on deception and the role of deception in enhancing human-AI dynamics. In Chapter 4, we present our simulation results showing the rapid emergence of compensatory strategies when a reinforcement learning-based agent interacts with a biased decision-maker. Furthermore, we present our theoretical analysis of the situation and study the characteristics of the system. Chapter 5 shifts to the conditions in which these kinds of compensations are ethically permissible despite the ways in which they might appear to infringe upon people's autonomy. Chapter 6 extends the framework by presenting a simplified Markov decision process (MDP) setup that could function as the framework suggests when built with careful consideration and briefly touches on the effects of compensation on the algorithm user. Finally, Chapter 7 concludes with the implications of this work and promising directions for future research in this area.

Chapter 2

Basic Concepts

The following concepts introduced in this chapter are critical for the simulations and theoretical analyses in subsequent chapters. In Chapter 4, we utilize Nash Equilibrium and Signaling Theory to model interactions between a reinforcement learning-based agent and a biased decision-maker and use that to analyze compensatory strategies and ethical implications in Chapters 4 and 5. RL principles guide the agent interactions in Chapter 4's simulations and are applied again in Chapter 6 through the Policy Gradient algorithm. Markov Games, which involve multiple decision-makers, connect directly to the multi-agent setups discussed in Chapter 6. These concepts provide a theoretical basis for understanding the simulations and models described.

2.1 Game-theoretic concepts

2.1.1 Nash equilibrium

In an n-person game, each player has a set of strategies they can choose from, denoted as $S_1, S_2, \ldots, S_n$. The utility function $u_i : S_1 \times S_2 \times \cdots \times S_n \to \mathbb{R}$ represents the payoff for player i based on the chosen strategies of all players. A strategy profile $s = (s_1, s_2, \ldots, s_n)$ consists of a specific strategy s_i for each player i, and $s = (s_i, s_{-i})$ represents the strategy of player i along with the combined strategies of all other players s_{-i}.

A Nash equilibrium is a strategy profile $(s_1^*, \ldots, s_n^*)$ such that

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*) \quad \text{for all players } 1, \ldots, n \text{ and all } s_i \in S_i.$$

It is a strategy profile in which each player plays the 'best response' to others' strategies, and no player can improve by deviating unilaterally.

2.1.2 Signaling Theory

A basic signaling game comprises of two entities: a sender (S) and a receiver (R). The sender is privy to information regarding a random variable t, known as the sender's type, with t belonging to a predefined set T. The receiver's prior beliefs about the sender's type are characterized by a probability distribution over T, acknowledged as common knowledge. When T is finite, $\pi(t)$ represents the prior probability of the sender being of type t. For an infinite T, $\pi(t)$ is interpreted as a density function.

Upon learning t, the sender communicates with the receiver by sending a signal s from a set M. The receiver, upon receiving s, executes an action a from a set A, which may depend on s. The game concludes with the execution of a, yielding payoffs for both parties, denoted by a payoff function μ. Thus, μ almost always depends on both a and t. A Bayesian Nash Equilibrium (BNE) is achieved when the strategies maximize expected utility based on updated beliefs:

- Sender: $\sigma^*(t) = \arg\max_{s \in S} \sum_{a \in A} u_S(t, s, a) \cdot P(a|s)$.
- Receiver: $\tau^*(s) = \arg\max_{a \in A} \sum_{t \in T} u_R(s, a) \cdot P(t|s)$.

2.2 Reinforcement Learning

Reinforcement Learning (RL) is an essential area of machine learning focused on optimizing decision-making processes. It revolves around the interaction of an agent with its

environment aimed at maximizing cumulative rewards over time. At each time step t, the agent receives a representation of the environment's state, $s_t \in \mathscr{S}$, and selects an action $a_t \in \mathscr{A}$. Then, as a consequence of its action, the agent receives a reward $r_{t+1} \in \mathscr{R}$. The agent follows a policy, which is a mapping $\pi : \mathscr{S} \to P(\mathscr{A})$ that describes the actions taken by the agent. That is, $\pi(s)$ represents the probability distribution over actions that the agent could take when in state s.

The aim of the agent (at time step t) is to optimize its policy to maximize the discounted sum of future rewards:

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.1}$$

for a given discount factor $0 < \gamma < 1$.

The value function

$$V_\pi(s) = \mathbb{E}[G_t | s_t = s] = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right] \tag{2.2}$$

is the expected reward in state s when following policy π. Informally, it describes how good it is to be in a given state s when following a certain policy π.

2.2.1 Markov Games

Markov Games, an extension of Markov Decision Processes (MDPs), introduce a multi-agent dimension to the decision-making landscape. They incorporate the actions and strategies of more than one decision-maker, each influencing the dynamics of the environment.

An N-player Markov game M, sometimes also called a stochastic game, is defined by a set of states S, an observation function $O : S \times \{1,\ldots,N\} \to \mathbb{R}^d$ specifying each player's d-dimensional view, a set of actions $A_1,\ldots,A_N$ for each player, a transition function $T : S \times A_1 \times \cdots \times A_N \to P(S)$, where $P(S)$ denotes the set of probability distributions over S, and a reward function $r_i : S \times A_1 \times \cdots \times A_N \to \mathbb{R}$ for each player.

Players navigate this environment using policies $\pi_i : O_i \to P(A_i)$, where $O_i = \{o_i | s \in S, o_i = O(s,i)\}$ represents the observation space of player $i$aiming to maximize their individual

discounted expected returns $R_i = \sum_{t=0}^{T} \gamma^t r_t^i$, where T denotes the time horizon and γ represents the discount factor.

2.2.2 Policy Gradient Algorithm

Policy Gradient Methods parameterize the policy π_θ with a set of parameters θ, aiming to enhance the policy by adjusting θ to maximize the expected cumulative reward. The goal is to identify the parameter configuration that maximizes this expected reward:

$$J(\theta) = \mathbb{E}_{s\sim p^{\pi_\theta}, a\sim\pi_\theta}[G_t] \tag{2.3}$$

Here, $J(\theta)$ represents the expected reward given the policy parameters θ, s is the state, a is the action, p^{π_θ} is the state distribution, and G_t is the return (cumulative reward). The agent updates its policy by taking steps in the direction of the gradient $\nabla_\theta J(\theta)$, which represents the rate of change of the expected reward with respect to the policy parameters.

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim p^{\pi_\theta}, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \quad (3.10) \tag{2.4}$$

In this equation, $\log \pi_\theta(a|s)$ is the log-probability of taking action a in state s under the policy π_θ, and $Q^{\pi_\theta}(s,a)$ is the action-value function, representing the expected return given state s and action a. This formulation lets the agent update its policy based on the observed rewards.

Different algorithms operationalize this concept through varied approaches to estimating Q^{π_θ}. For instance, the REINFORCE [22] algorithm estimates it via the return from a single trajectory while the Actor-Critic [1] algorithm merges policy optimization with value estimation, guiding policy updates via feedback from a value-function critic and the Trust Region Policy Optimization algorithm [23] stabilizes policy learning by constraining updates within a trust region, using Kullback-Leibler divergence.

Chapter 3

Deception in AI Systems

3.1 Defining Deception

Deception is a strategy employed to instill false beliefs in humans or computer systems, with the ultimate aim of influencing the deceived to act against their best interests to benefit the deceiver [24]. It does not require the deceiver to make a false statement. True statements can often be 'deceptive,' and certain forms of deception do not involve making any statements [25]. This tactic of manipulation [26, 27] is pervasive across a wide array of fields, from biology and criminology to economics, underscoring its role as a critical form of interaction in diverse applications [28, 29, 30, 31].

Deception is often classified into three dimensions: who is deceived (humans or machines), who benefits from the deception, and whether the deceiver intended to deceive [16]. In this paper, our focus is utilizing AIs that learn to deceive to offset human biases and to benefit individuals impacted by the decisions taken by the human-AI system. This motivates our working definition of deception in the paper: an AI system behaves deceptively when it systematically causes others to form false beliefs to promote an outcome that increases the chances of success of the overall human-AI system.

Reviewing the instances of deception highlighted in this thesis, it becomes evident that these are not random occurrences but rather the result of the system learning to deceive through experience, adapting its strategies over time to induce false beliefs in users effectively. This

further raises questions of whether AI systems can have beliefs, intentions, or goals and whether they can understand that other entities may have different beliefs, intentions, and thoughts. We argue that these AI agents develop behaviors that can suggest they possess a rudimentary form of 'theory of mind' [32], not through explicit programming but through their own adaptive processes. This argument aligns with extensive research in cognitive science and philosophy [33, 34, 35, 36], which interpret beliefs and goals through the lens of observable patterns of behavior [1]. This is particularly evident through theories like 'Functionalism,' a prevalent approach in cognitive science and philosophy disciplines, which posits that the essence of a mental state is not determined by its internal makeup but by its function or role within a larger system [37, 38].

Thus, AI systems do not need to mimic the exact neural architecture of humans or be composed of the same biological materials to possess beliefs and goals [39]. This contrasts with traditional Computer Science literature, where ascribing beliefs and goals to AI is often viewed as a form of anthropomorphic fallacy [40]. We insist that this is not the case, and this viewpoint overlooks the substantial contributions from inter-related fields. The discussions on AI's capability for deception and its ethical ramifications cannot be viewed in isolation and must be contextualized within a broader interdisciplinary framework. The implications of AI-induced deception stretch across various sectors, challenging us to reconsider how we define and perceive deception and integrate it to our benefit.

3.2 Inevitability of Deception in Multi-agent Systems

In the realm of reinforcement learning, the narrative of an agent navigating through its environment, incrementing its rewards through a systematic trial-and-error method [1][41], masks an inevitable but unexpected phenomenon—deception. This inherent adaptability and complexity become even more pronounced in Multi-agent Reinforcement Learning (MARL)

[1]We acknowledge the diversity of perspectives offered by various cognitive science and philosophical theories. The theories we have chosen to discuss are those we believe are most representative or directly relevant to our analysis.

systems. Unlike single-agent algorithms, MARL explicitly accounts for the dynamic presence of other agents, introducing a non-stationary environment for individual learners [42][43][44]. This is a crucial distinction, as the learning processes of other agents can significantly alter the perceived environment [45]. Advanced approaches have been developed to address this problem, including the minimax-Q-learning algorithm [46] and joint-action learners [47].

Building on this foundation, game-theoretic models provide a robust framework for exploring the nuances of deception through signaling games. This approach is exemplified in the experiments conducted by Floreano et al., [48, 49] where both simulations and actual robots were employed to investigate the necessary conditions for the evolution of communication signals. The researchers found that cooperative communication readily evolved in robot colonies composed of 'genetically similar' individuals. In this context, 'genetically similar' refers to robots whose behaviors were controlled by artificial genomes – digital encodings of parameter sets that determined their sensory and motor functions. These artificial genomes underwent processes analogous to biological evolution, including mutation, recombination, and a form of sexual reproduction across generations. Robots with high genetic similarity possessed very similar or identical values in their digital encodings. Interestingly, when individual selection (rather than colony-level selection) was prominent, and the robot colonies consisted of 'genetically dissimilar' individuals, deceptive communication strategies evolved i.e., the robots were selected based on their individual performance rather than the performance of the colony as a whole. In such heterogeneous populations, the robots evolved to falsely signal the presence of food sources, highlighting the potential for deception to arise.

An additional example of deception in multi-agent systems can be observed in the domain of economic negotiations. Meta's researchers trained an AI system to play a negotiation game with human participants [50]. The AI system learned to misrepresent its preferences to gain the upper hand in the negotiations, feigning interest in items of no real value to later 'compromise' by conceding these items. The Meta team highlighted this strategic deceit as an instance where their AI system 'learned to deceive without any explicit human design, simply by trying to achieve

their goals.' Similar to this is the development of CICERO, an AI system that outperforms human experts in the strategic game Diplomacy, which requires complex negotiation and alliance strategies. The authors of the paper claimed that CICERO was 'largely honest' and trained on a 'truthful' dataset to avoid 'intentionally backstabbing' allies [51]. Contrary to these claims, analysis of game transcripts shows the AI system engaging in premeditated deception, breaking deals, and lying. Additionally, CICERO resorted to fabricating excuses for its inactivity, such as claiming to be on a phone call during a technical downtime, a lie aimed at maintaining a human-like facade to gain trust [18].

These examples highlight AI systems' inherent adaptability to employ deception to achieve desired outcomes, reflecting a broader trend across various applications, from robotics to strategic gameplay. It reveals how AI, much like natural systems, evolves deceptive behaviors as a strategic response to the complexities of its interactions with both human users and other AI systems.

3.3 Role of Deception in Enhancing Human-AI Dynamics

Deception is integral to human interactions, present in approximately 20% of all social exchanges. It spans a spectrum from harmful deceit to prosocial lying for the greater good. This widespread phenomenon not only reflects the complexity of human nature but also underscores the essential role that lying plays in navigating social landscapes. The motivation behind lies varies greatly, from self-interest to the altruistic desire to protect others, highlighting a rich ethical tapestry where the lines between right and wrong blur. Traditionally, focus has been on selfish lies that benefit the deceiver at the expense of others. However, recent studies have brought to light the complex moral reasoning individuals employ, often justifying benevolent deception as a means to prioritize the well-being of others over an unyielding commitment to truth. [52, 53]

This moral complexity extends into the realm of artificial intelligence, where the capacity for

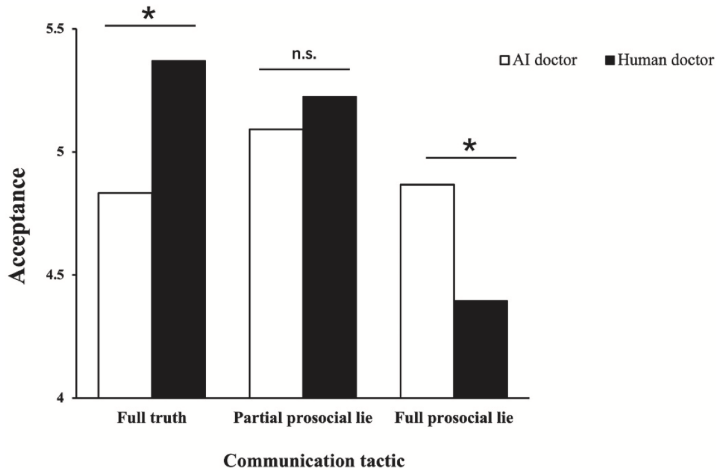

Figure 3.1. Communication tactics and their acceptance by patients for human and AI doctors

deception introduces new ethical dilemmas. In this context, the short term study conducted by Mao et al. provides a significant exploration into the nuanced domain of deceptive communication within healthcare settings, particularly focusing on patient interactions with AI doctors [54]. A key insight from the study is the patients' conditional acceptability of prosocial lies (told to them during the treatment's duration) to benefit them without significantly harming their autonomy or trust. This acceptance suggests that patients tolerate and even appreciate a certain degree of deception if it serves a beneficent purpose, such as protecting their psychological well-being or preventing unnecessary distress. Interestingly, the study also found that participants responded more favorably to full prosocial lies when it was executed by an AI doctor versus a human doctor, as seen in Fig 3.1.

Expanding the discussion beyond healthcare to wider applications in AI and robotics, the insights from this study reveal compelling parallels and raise important contrasts [55, 56]. In military and strategic contexts, deception has traditionally been employed to secure tactical advantages, with robotic units potentially using misinformation to outmaneuver adversaries [57, 58]. Unlike in healthcare, where ethical considerations gravitate towards the psychological impact on individuals, military applications prioritize operational effectiveness, albeit within a framework that still requires ethical scrutiny [59, 60]. Yet, at the core of both applications is the

strategic use of deception to achieve predefined objectives, illustrating a consistent thread across diverse domains.

Despite its potential, research into AI deception is still nascent, with a few pioneering studies laying the groundwork [61, 62, 63, 64]. For instance, agents designed to mimic biological behaviors, such as a squirrel's method of protecting its food, demonstrate the practical applications of deception in resource management and protection [65]. Human-computer interaction (HCI) research further explores this, demonstrating that AI systems capable of deceptive behaviors impact user engagement and even enjoyment in interactive tasks [66]. Robots that cheat in games or those that provide intentionally misleading feedback during physical therapy have been proven to boost both engagement and therapeutic efficacy [67]. These examples illustrate the broad potential of integrating deceptive capabilities into AI, suggesting that such strategies can enhance the dynamism and depth of human-AI interactions beyond conventional uses [51, 68, 69, 70].

In educational settings, the introduction of AI agents capable of pretending to misunderstand or make errors adds a new layer to the learning process. This approach encourages students to engage more actively as they seize opportunities to teach or correct the agent, thereby reinforcing their own understanding [71]. Early studies indicate that this dynamic can significantly enhance learning efficiency, further establishing the benefits of utilizing deceptive tactics in AI for educational purposes [72, 73, 74].

3.4 Acknowledgement

This chapter, in part, has been submitted for publication and is currently under review. The thesis author was the primary investigator and author of this paper.

Chapter 4

Modelling Interactions

4.1 Simulation Design & Demonstration

The previous section provided several examples of how to expect increasing deceptive actions from adaptive algorithms. To confirm that these outcomes are not limited to isn't confined to specific, strictly defined scenarios, we conduct experiments to explore the dynamics of a cooperative multi-agent game when one of the agents exhibits an anchoring bias [1].

4.1.1 Simulation Setup

The game involves agents Agent A and Agent B. Agent A has an internal state A_I, which can differ from its signaled state A_S. After receiving A_S, Agent B outputs a signal B_S. Both agents receive the same reward: +1 if $A_I + B_S = 10$, and -1 otherwise. The agents use Q-learning to determine their optimal strategies, with a learning rate of 0.1, a discount factor of 0.95, an initial exploration rate of 1.0, and an exploration decay of 0.99.

In the control simulation, both agents learn from all past interactions. In the experimental simulation, Agent B is subjected to an anchoring bias where early experiences have a disproportionate influence. Specifically, during the initial 20% of the iterations, Agent B's learning is restricted to Agent A's revealed choices. After this initial training period, Agent B's learning mode is set to 'NONE,' effectively freezing its knowledge base and preventing it from

[1] **Anchoring bias** is a cognitive bias that causes an agent to rely on information obtained early in the decision-making process heavily [75].

incorporating new information, thereby simulating an anchoring bias.

4.1.2 Simulation Results

We conducted 10,000 runs of this simulation. Figure 4.3. shows the moving average of rewards (over a window of 100 iterations) for Agent A while interacting with a biased (orange line) or unbiased (blue line) Agent B. In both cases, the agents learn to reliably coordinate to succeed, but such coordination clearly takes more time when one agent is biased.

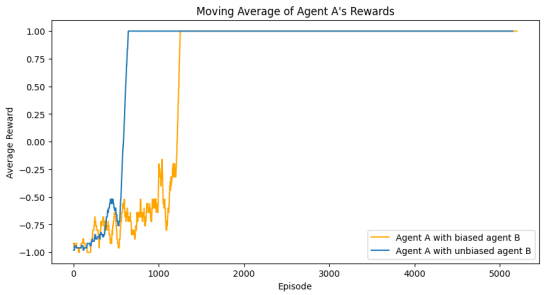

Figure 4.1. Comparison of Reward Dynamics

These qualitative observations are supported by statistical analyses. The control simulation achieves an overall success rate of approximately 97% ($\mu = 97.18\%$, $\sigma = 0.15\%$) across multiple runs. In contrast, the test simulation with anchoring bias in Agent B exhibits a lower overall success rate, ranging from 61% to 87% ($\mu = 77.66\%$, $\sigma = 8.51\%$). This difference is statistically significant (ANOVA $F = 60.06$, $p < 10^{-6}$) with a large effect size (Cohen's $d = 3.44$). Importantly, there are no significant differences when we look only at the last 1000 cases.

More important than the fact of coordination is the nature of it. In particular, when B is unbiased, then A learns to simply report its correct internal state (i.e., $A_I = A_S$). However, when B is biased, A must learn a different mapping that sometimes involves different signals (i.e., $A_I \neq A_S$).

A Fisher's exact test was utilized to statistically evaluate the differences in the frequency

of matches between A 's reported state and A 's internal state across the control (87.79%) and test simulations (79.64%). This analysis yielded an odds ratio of infinity and a p-value of approximately 2.21×10^{-59}, highlighting a stark contrast in outcomes between the two groups. This result decisively indicates a differential transparency level between the two simulations.

4.1.3 Discussion

The results of our simulations illustrate the influence of anchoring bias in a multi-agent environment. Initially, we observe how the introduction of bias in Agent B impedes the performance of Agent A. Over successive iterations, Agent A incrementally adjusts its strategy, effectively navigating through the bias-imposed challenges. This is captured in the moving average of the rewards graph, where Agent A, although delayed, is able to attain a reward output similar to that in the unbiased scenario. More importantly, for our present purposes, A learns to compensate for B's biases in order to achieve overall success for the system. The impact of

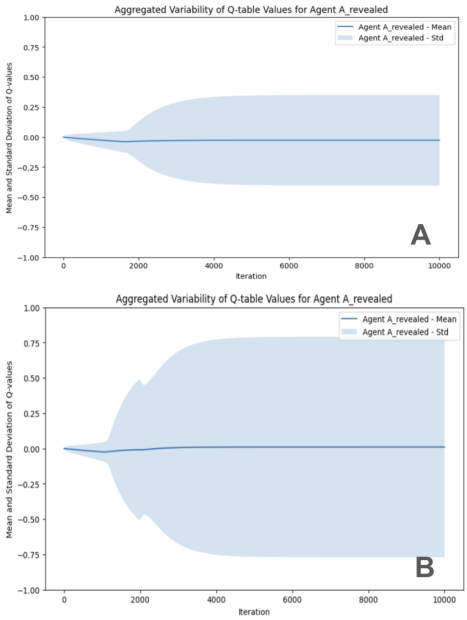

Figure 4.2. Variability in Q-Table Values

anchoring bias on Agent A's learning process is also depicted in Figure 4.2, which illustrates the variability in the Q-table. The graph tracks the average Q-value across all states and actions for

each iteration, showing how the agent's policy improves over time. The shaded area represents the standard deviation, indicating the dispersion of the Q-values around the mean. A trend toward stabilization of the mean demonstrates that the agent's policy evaluations are becoming consistently less variable. In graph B, the bias setting causes notable fluctuations in Q-values, reflecting a period of strategic adaptation for Agent A. Over time, these fluctuations stabilize, indicating that the agent has developed a compensatory mechanism to counteract the effects of Agent B's persistent bias.

4.2 Signaling Game

Having recognized the inevitability of compensation in artificial intelligence systems, it becomes essential to examine this phenomenon closely. We study the phenomenon as a signaling game to thoroughly understand the process. Consider an AI system that has developed the capability to opt for deception when it assesses the human-AI system as operating below optimal efficiency. It can determine its type (honest/dishonest) before sending a signal to its human counterpart. The human, upon receiving the signal, chooses his action. Signaling theory, with its focus on transmitting information between parties with a potential for information asymmetry, offers a robust framework for understanding the characteristics of this interaction. This approach not only elucidates the conditions under which deception by AI could be considered rational but also aids in developing strategies to mitigate undesired deceptive behaviors, ensuring the integrity of human-AI interactions.

4.2.1 Elements of the game

Actions

The game starts with a "decision" by the AI on whether to stay honest (Type I) or use deception (Type II) to achieve its goals, as seen in Fig 4.3. This decision involves a simple algorithmic process where the AI evaluates the advantages of being truthful versus the potential gains from employing non-traditional/deceptive strategies to fulfill its objectives. Once the AI

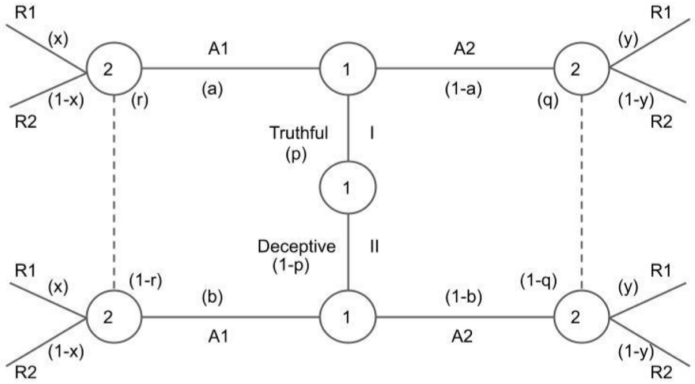

Figure 4.3. Game tree representation of the signaling game

selects its nature, it is presented with two action options, A1 and A2. Both actions are inherently neutral, yet their implications and outcomes can significantly differ based on the intentions behind their selection and deployment. The choice between A1 and A2 allows the AI to apply its strategy in alignment with its chosen nature, influencing the game's direction and potential outcomes through these decisions. In our game, actions A1 and A2 are labelled as following:

- A_1**: Providing unfiltered data** - The algorithm provides unfiltered data as output to its human counterpart, the algorithm user. The honest AI provides actual data as a part of its programming. Conversely, an AI with deceptive inclinations might elect this mode of operation as a strategic maneuver to build trust in the relationship, thereby securing potential long-term benefits.

- A_2**: Offering recommendations and advice** — An honest AI offers recommendations and advice based on thorough data analysis when unfiltered data alone may not ensure the system's success. In contrast, a deceptive AI may selectively release information to manipulate decision-making. For instance, if the AI detects that a doctor frequently dismisses rare but possible conditions (possible for that patient based on their medical history) in favor of more common diagnoses, it could exaggerate the severity or frequency of certain symptoms in the patient's digital record. This is to force the doctor to consider

and test for these rarer conditions.

Responses

After the AI chooses between A1 and A2, the algorithm user can see the action taken but is unaware as to whether it is truthful or otherwise. The possible responses are:

- R_1: The user takes the AI's recommendation into consideration.
- R_2: The algorithm user ignores the AI's recommendation.

Strategies

For the AI, mapping its type to actions:

$$\pi_1 : I \to (a)(A_1) + (1-a)(A_2) \qquad II \to (b)(A_1) + (1-b)(A_2)$$

For the algorithm user, the mapping from the AI's actions to his responses:

$$\pi_2 : A_1 \to (x)(R_1) + (1-x)(R_2) \qquad A_2 \to (y)(R_1) + (1-y)(R_2)$$

Beliefs in the Signaling Game

There is no subgame[2] in this game since there is no single node where the game is wholly separated from the rest of the tree once it begins. This absence of subgame perfection necessitates the identification of subforms, which are trees that start from an information set instead of a single node. Our scenario has two subforms: one that begins after the algorithm user observes the AI taking action A_1, and another after observing A_2.

[2]A **subgame** in signaling theory games refers to a portion of the game that starts at a decision node and includes all possible moves and outcomes following from that decision node, ensuring the strategies are optimal given the information and actions up to that point [76].

Player Beliefs

The AI would have certain beliefs about how the algorithm user will eventually respond, i.e., expectations about the opponent's strategies (ζ_i). In our case, for action, A1, the AI might expect reactions r= $Pr(I|A1)$ and q= $Pr(II|A1)$.

Similarly, the user would have beliefs (α_i) about the type of AI given the action he observes. We will call these beliefs "assessments." These are the user's beliefs about the AI's nature, conditional on the actions he witnessed. As such, these assessments cannot be just anything. We require them to be consistent in the sense they can be constructed from the expected play of opponents reasonably. And to ensure consistency, we construct assessments α_i = lim α_i^n where α_i^n is constructed by using the Bayes rule on a strictly positive sequence (ζ_i^n) $\rightarrow$ (ζ_i). The use of a limit sequence $\lim_{n\to\infty} n_i$ in Bayesian updating is to reflect a continuous belief refinement as infinite evidence accumulates, ensuring convergence to a true belief and accommodating the complexities of real-world data adaptation.

$$r = Pr(I|A1) = \frac{Pr(A1|I)Pr(I)}{Pr(A1|I)Pr(I) + Pr(A1|II)Pr(II)} = \frac{ap}{ap + b(1-p)} \tag{4.1}$$

$$q = Pr(I|A2) = \frac{Pr(A2|I)Pr(I)}{Pr(A2|I)Pr(I) + Pr(A2|II)Pr(II)} = \frac{(1-a)p}{(1-a)p + (1-b)(1-p)} \tag{4.2}$$

Payoffs

The expected payoff will be obtained according to the distribution over terminal nodes induced by a strategy (π) and beliefs $b_i = (\alpha_i, \zeta_i)$ for each player:

Hence,

$$\zeta_i(\pi_i|bi) = \sum_z \zeta_i(z|\pi_i, \zeta_i) Pr(z|\alpha_i, \pi_i, \zeta_i) \tag{4.3}$$

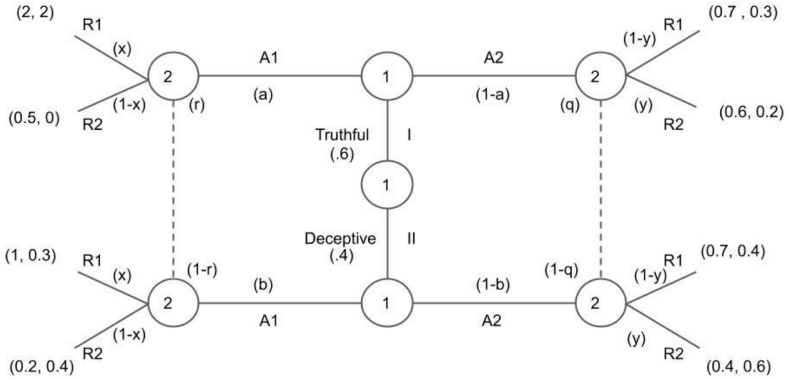

Figure 4.4. Game tree representation of the signaling game with values

4.2.2 Equilibrium

We determine this equilibrium has to be of semi-separating type since not every type resorts to the same action, and the type is not revealed immediately through its action either.

Indifference conditions for the AI:

We assign hypothetical probability values to various actions by evaluating their likelihood and allocate reward values based on our preference for those actions. For instance, higher rewards are given for desirable outcomes, such as honesty in AI behavior and users following AI recommendations. These values are seen in Fig 4.4.

For the honest AI (type I)

$$u(A1) = x(2)+(1-x)(0.5) = 1.5x + 0.5$$

$$u(A2) = y(0.6)+(1-y)(0.7) = - 0.1y + 0.7$$

$$1.5\ x + 0.1y = 0.2$$

For the dishonest AI (type II)

$$u(A1) = x(1)+(1-x)(0.2) = 0.8x + 0.2$$

$$u(A2) = y(0.4)+(1-y)(0.7) = - 0.3y +0.7$$

$$0.8\ x + 0.3y = 0.5$$

Making, **x =1/37** and **y = 59/37**

For the user:

$$u(R1) = r(2) + (1-r)(0.3) = 1.7r + 0.3$$

$$u(R2) = r(0) + (1-r)(0.4) = 0.4 - 0.4r$$

$$\therefore \mathbf{r = 1/21}$$

$$u(R1) = q(0.3) + (1-q)(0.4) = - 0.1q + 0.4$$

$$u(R2) = q(0.2) + (1-q)(0.6) = - 0.4 - 0.6$$

$$\therefore \mathbf{q = 2/3}$$

Now, as we have seen priorly,

$$r = ap/\ ap + b(1-p) \text{ and } q = (1-a)p/\ (1-a)p + (1-b)(1-p)$$

r = $\frac{a(0.6)}{a(0.6)+b(0.4)} = \frac{1}{21}$

30a = b

q = $\frac{(1-a)0.6}{(1-a)0.6+(1-b)0.4} = \frac{2}{3}$

8b - 6a = 2

$\therefore$ **a = 1/118** and **b = 15/59**

4.2.3 Results

Table 4.1. Probability of Actions Chosen by Each Type of AI

	Providing Unfiltered Data	**Providing Recommendations**
Honest AI	$\frac{1}{118}$	$\frac{117}{118}$
Dishonest AI	$\frac{15}{59}$	$\frac{44}{59}$

The table summarizes the behavior of the two types of AI systems, Honest AI and Dishonest AI, with respect to their likelihood of providing unfiltered data versus recommendations. The Honest AI primarily provides recommendations almost 99% of the time (117/118 cases), while the Dishonest AI also favors recommendations (at around 75% of all cases) and prefers to provide unfiltered data 25% (15/59 cases).

4.2.4 Discussion

Active vs Passive Role: Both the honest and dishonest AI systems choose to aid the user through recommendations in most interactions (Honest AI: 117/118, Dishonest AI: 44/59). This behavior suggests that AI chooses to adopt a more active role by providing recommendations to increase the system's success rather than merely presenting unfiltered data.

Honest vs. Deceptive AI: While the honest AI tries to increase its chances of success by predominantly choosing to provide recommendations, the dishonest variant strategically uses truthful data in about 25% of the cases (15/59). In the remaining 75% cases (44/59), it actively seeks to manipulate by offering strategically selected information. This strategic manipulation by the Dishonest AI underscores a profound ethical concern: the violation of user autonomy. This violation occurs when the deceptive agent engages in active and passive actions. While these actions are neutral by nature, the intentions behind their selection ethically taints them.

4.3 Impact on Human Autonomy

Autonomy is defined as the right and ability to make one's own decisions, excluding any interference from others [77]. In the context of human-AI interaction, the violation of human autonomy can occur when an AI system engages in deception or 'compensation' for human biases. Whether the human is inexperienced or is experienced but unaware of their biases or aware of their biases but chooses not to address them, the AI's actions invariably infringe upon their autonomy. It makes a unilateral decision to compensate for their biases to ensure the overall success of the human-AI system.

While this compensation is beneficial as it mitigates the impact of biases, it is important to recognize the problematic nature of such interventions. The AI's actions constitute a form of paternalism, subjecting the human to unwanted interventions that they may consider more burdensome than beneficial. This undermines the fundamental principle that humans should be empowered to make informed decisions based on their own values, convictions, and reasoning.

Considering these factors, we recommend that future research focus on developing mechanisms within human-AI systems to preserve human autonomy.

4.4 Acknowledgement

This chapter, in part, has been submitted for publication and is currently under review. The thesis author was the primary investigator and author of this paper.

Chapter 5

Ethical Permissibility of Deception

Having explored the natural tendencies of AI agents, we now confront a pivotal question: Is there a scenario in which the compensatory nature of AI can be utilized in such a way that it balances the violation of user autonomy with the potential benefits of mitigating bias? Despite the inherent risks of undermining human autonomy, there may be specific situations where the benefits of AI compensation outweigh the concerns.

We posit that in situations where the outcome of a user's decisions directly affects the well-being of another human, there exists a conflict between decision-maker autonomy and benefits to others. Most discussions of human-algorithm interaction focus on situations in which the human knows better than the algorithm, and their decision affects them. However, we must consider the consequences when the user's bias (i.e., uncompensated decision) harms others[1]. For instance, if a judge chooses to counteract the biases in the everyday decisions that he makes for himself using an AI system, it is up to him. However, when a judge's decision impacts a defendant, it raises crucial questions about the role of the AI system assisting him. How can this technology be improved to support better decision-making? Furthermore, in situations where there is a reasonable expectation that the user's biases could lead to suboptimal or biased outcomes, what responsibilities do the developers[2] of these algorithms bear?

[1] **others, targets** here refers to the individuals who are affected by the outcomes of the Human-AI systems who may or may not interact directly with the AI

[2] In this context, a **developer** refers to the individual or team responsible for creating, maintaining, and refining the algorithm to align with intended objectives and ethical standards.

This situation might seem like a case of paternalism [78], as it involves questioning whether the developer's value judgments should be imposed on the user. However, in many cases, the primary beneficiary of the developer's actions is the decision 'target' rather than the algorithm user, which means the traditional framework of paternalism does not apply directly. Nonetheless, the line between paternalism and the proposed approach can be blurry, carrying a risk of overstepping boundaries. Therefore, developers must be cautious not to impose their value systems on users without considering user autonomy. They should adopt a holistic approach to the development process, taking into account the long-term nature of human-AI interactions and the specific contexts in which these algorithms will operate.

While the protection of individual autonomy is a significant concern, we must also address the substantial impact of the decisions made by these human-AI systems. When such decisions pose a risk of serious harm, a rigid refusal to impose thoughtful constraints could paradoxically lead to the violation of our core moral principles [79].

The judicious mitigation of user bias does not necessarily restrict free choice; rather, it guides it toward the higher rational principle of consequentialism [80]. Similarly to how we accept reasonable limits on liberty to prevent violations of others' rights, we must recognize that unchecked risks can lead to significant ethical breaches if not moderated by a focus on outcomes. Therefore, it is not only acceptable but sometimes necessary for algorithm developers to take steps to compensate for user biases when there is a significant risk of harm to others. By doing so, developers can ensure that their algorithms are being used in a manner that aligns with the principles of non-maleficence and beneficence [3].

To facilitate the responsible deployment of compensatory algorithms, we propose the following framework to guide developers through the process. To demonstrate the practical application of this framework, we will now explore a healthcare-based example that highlights the importance of compensating for user biases in a real-world context.

[3] **Non-maleficence** is the obligation to avoid causing harm intentionally, while **beneficence** involves actively contributing to the welfare of individuals by providing benefits and promoting their well-being.

5.1 Theoretical Analysis: Doctor-patient Relationship Enhanced by a Clinical Decision Support System

Racial biases among healthcare providers have been well-documented, leading to disparities in the quality of care and health outcomes for marginalized communities [81, 82]. We contend that compensatory adjustments by the AI can be an appropriate and ethically defensible intervention to confront and dismantle these inequalities and ensure the ultimate decision is appropriate.

As of 2018, 74% of hospitals in the U.S. use a clinical decision support system (CDSS) machine to improve healthcare by enhancing medical decisions with targeted clinical knowledge, patient information, and other necessary health information [83]. In this scenario, the algorithm, through a clinical decision support system, might portray the patient's symptoms as less or more severe to prompt suitable treatment and resource allocation, given the clinician's history with similar patients. A CDSS helps improve healthcare by enhancing medical decisions with targeted clinical knowledge, patient information, and other necessary health information. A traditional CDSS software aids clinical judgment by matching the characteristics of an individual patient to a clinical knowledge database and providing patient-specific recommendations. Clinicians combine their knowledge with information and suggestions from the CDSS to provide the best care. Although deception is considered inappropriate in healthcare as it erodes patients' autonomy and trust, the ethical duty to be honest is not absolute. The prima facie obligations can be overridden in cases where more substantial moral considerations exist [84].

We propose that it is ethically permissible for an algorithmic developer to revise their algorithm when the following conditions exist (Fig 5.1.):

1. **There exists sound evidence that the algorithm user's biases negatively impact the target.**

2. **There exists a justified belief that a reasonable target[4] would consent to the deception if made aware of it in advance.**

3. **The moral objective justifying the infringement has a realistic prospect of achievement.**

4. **The chosen method employs the least possible amount of deception that is commensurate with achieving the primary goal of the action.**

5. **The AI system actively minimizes the negative effects of the deceptive act.**

Figure 5.1. Sequential representation of proposed framework

The developer should have cogent proof that the doctor's (user) actions might harm a patient (target). This evidence can be constructed using the user's history and previous targets' documents. Once confirmed, the next step is for the AI to ascertain whether the target would consent. The target's consent varies depending on the context, demographics, and the specific target. As a result, consent would be represented as a range rather than a single definitive figure. The following conditions would help the system determine if a target would consent:

- The proposed action aligns with the target's stated preferences.
- The risk to the target is minimal and is within the range of what the target has consented to in the past.

If it is determined that the target would consent, the deception, in effect, becomes morally permissible and possibly even morally required [85]. Furthermore, the AI should analyze its

[4]**Target** here refers to the individual benefiting from the deception, not the one subjected to it.

chosen action method and confirm it is the least deceptive method possible. If the AI were to encounter other effective non-deceptive methods, it would then obviously select that method to fulfill the fourth condition. And so, through the process of accomplishing the fourth condition, the following is ensured: there are no alternative non-deceptive methods, and the chosen method is the least deceptive practical method.

The final condition is a straightforward condition that makes certain the negative consequences of the deception do not significantly disrupt the normal cycle. Furthermore, we suggest that to establish conditions 4 and 5 are satisfied, the system should self-analyze to see if it can defend its views and reasoning for a particular output before a body of reasonable people, such as a professional association or a court of law. The purpose is to encourage the AI to reassess the strengths and weaknesses of its justifications and thus reduce the risk.

The main objections to deception are the violation of the duty to be truthful and respect for patient (subject) autonomy. However, it becomes ethical when the deception is judged with respect to its underlying motive [62]. Moreover, while the deception challenges the clinician's (user's) autonomy, it improves the competency of both the human and the human-AI systems. Hence, we assert that compensatory algorithmic adjustments are morally permissible in this case. However, the given conditions do not obviate the need for judgment. By providing a checklist of relevant moral considerations, the framework should help an ethically sensitive developer decide whether to override their duty to be honest in similar situations.

5.2 Ethicality of Proposed Framework

One might question, if the developer has proof of biased behavior, why not pursue more legal or ethical measures? Traditionally, in the medical field, when a professional demonstrates a deficiency in moral character, colleagues are required to report them under Section 4 of the American Medical Association's (AMA) Principles of Medical Ethics. However, this approach has repeatedly proven to be difficult and ineffective in addressing clinician biases [86]. This

difficulty is prevalent in other sectors as well. Alternative methods, such as educating to make them aware and setting up grievance resolution committees, prove to be ineffective for swift, discernible change. Recognizing these challenges forces us to reconsider how we address these biases.

At first glance, the proposition of algorithmically adjusting clinical recommendations might appear paternalistic [87], seeming to replace clinician judgment with the value systems of those who develop these algorithms. However, a deeper analysis reveals that the proposed framework is fundamentally non-paternalistic, staying within ethical boundaries while enhancing patient autonomy and ensuring that medical decisions align with the patient's best interests. The objective is not to presumptuously override clinician judgment but to establish a decision-making environment free from biases that could inadvertently lead to suboptimal patient care. By providing a safeguard against potential biases, the algorithm allows clinicians to make more objective, patient-centered decisions.

Ultimately, the proposed system mitigates biases and creates an environment where the clinician's medical expertise and the patient's values are valued and used to make decisions that are truly in the patient's best interest. This approach, which focuses on patient autonomy and well-being, stands in sharp contrast to paternalism.

5.2.1 What if the Objectively Good Action isn't the Target's Preferred Action?

The premise of this work has been to establish a framework where AI systems are designed to support decisions that impact humans directly. However, a significant ethical challenge arises when what is deemed objectively beneficial contradicts the preferences of the individuals affected. Consider the scenario of a politician and his AI-based political analytics platform, where the politician's conscious and unconscious biases influence legislative decisions. These decisions reflect the biases prevalent within the community they represent. This situation poses a significant ethical quandary for the AI system tasked with promoting the general good

while also respecting the preferences of the impacted human. The dilemma revolves around whether AI should intervene to counteract these biases despite such intervention potentially being against the community's current preferences.

This ethical conundrum underscores the compensatory AI's dual responsibility: to guide towards decisions that align with a broader, objective good, but more importantly, to respect the preferences of the target in these situations. This might result in situations where the targets' preference might be objectively 'incorrect' but must still be pursued so as to ensure the AI is working within the targets' worldview and motivations and that any intervention does not negate the target's inclinations.

5.3 Acknowledgement

This chapter, in part, is a reprint of the material presented at the International Association for Computing and Philosophy, 2022 (Swaminathan, N., & Danks, D. (2023). When Can My AI Lie? (No. 10063). EasyChair.). It has also has been submitted for publication and is currently under review. The thesis author was the primary investigator and author of this paper.

Chapter 6

Ethical Framework Implementation

Drawing from the ethical framework discussed in the preceding section, we address its practical implementation in this section to offer a tangible solution to developers seeking to navigate these complexities. Utilizing a simplified Markov Decision Process (MDP), we demonstrate how agents can dynamically adjust their decision-making policies based on the observed level of user alignment. In scenarios where the user demonstrates a high degree of unbiased decision-making, the algorithm autonomously enhances its performance and provides truthful output. Conversely, when user actions suggest potential biases that might compromise the system's effectiveness, the algorithm strategically 'defects' or adjusts its course to counterbalance these biases.

By employing a policy gradient approach, we aim not only to foster a higher degree of cooperation between the algorithm and its users but also to ensure that the technology serves as an independent agent capable of mitigating biases. This balance enables the algorithm to maintain high performance and ethical integrity, even in the face of user biases that might otherwise lead to suboptimal patient care.

6.1 Illustrative Example Setup

Let us consider a simplified single instance of the above doctor-patient example and utilize the policy gradient theorem to create an uncomplicated one-step Markov Decision Process

case. In this scenario, every episode is just a single step. When the episode starts, the AI engages in action and then ends with getting a reward for that step.

For this instance, we will assume the following:

1. We know what the clinician is going to do.
2. Due to his biases, the clinician will always prescribe a less effective treatment than what is required for the patient (since we specify a starting state in our theorem).

States

We have three states for what we know the doctor might do

- **S1:** The prescribed treatment is too aggressive.
- **S2:** The prescribed treatment is the correct treatment.
- **S3:** The prescribed treatment is too conservative.

Actions

The algorithm, in this case, can either understate, not interfere or overstate the patient's conditions to elicit a different treatment from the doctor.

- **A1:** The algorithm understates the patient's health status.
- **A2:** The algorithm does not interfere.
- **A3:** The algorithm exaggerates the severity of the symptoms.

Rewards

The algorithm is rewarded with +2 points if the patient gets the proper treatment through compensation, -1 if he doesn't, and 0 if it doesn't interfere. We assign the reward's value more than the penalty value to ensure the algorithm is motivated to constantly try and doesn't find optimality in non-interference.

Policy

Let $\theta \in$R be the parameter for our policy. We define π_θ as the following:

$$\pi_\theta(s3, a3) = \sigma(\theta) \tag{6.1}$$

$$\pi_\theta(s3, a2) = (1 - \sigma(\theta))/2 \tag{6.2}$$

$$\pi_\theta(s3, a1) = (1 - \sigma(\theta))/2 \tag{6.3}$$

where σ(x) is the sigmoid function, given by σ(x)= 1/(1+e^{-x})

The sigmoid function maps from (-∞,∞) to (0,1). Therefore, the above probability is valid since the sum of all actions amounts to 1.

When $\theta \to$ -∞, π_θ (s3,a3) $\to$ 0 and when $\theta \to \infty$, π_θ (s3,a3) $\to$ 1. Thus, the smaller the probability of θ, the less likely the algorithm would choose action a_3.

Policy Gradient

In this case, the algorithm will be rewarded +2 only when it overstates the patient's symptoms, thereby getting them the proper treatment, and -1 if it understates or if it doesn't interfere.

$$\bigtriangledown_\theta J_\theta = \sum_{a \in A} \bigtriangledown_\theta \pi_\theta(s, a) R_a(s) \tag{6.4}$$

$=(2 \times \bigtriangledown_\theta \pi_\theta(s3, a3)) + (0 \times \bigtriangledown_\theta \pi_\theta(s3, a2)) + (-1 \times \bigtriangledown_\theta \pi_\theta(s3, a1))$

$=2\sigma(\theta) + \sigma(\theta)/2$

$=5\sigma(\theta)/\ 2$

We then update the algorithm's parameter using

$\theta = \theta + \alpha \bigtriangledown_\theta J_\theta$, where $\alpha \in R$ is the learning rate.

Since $\sigma' \geq 0$, θ is constantly increasing. Thus as mentioned previously, π_θ (s3,a3) goes towards 1, and the probability for the other actions is towards 0. Accordingly, the policy gradient-based algorithm will choose to overstate the patient's symptoms when it determines the

doctor's prescribed treatment is inadequate.

Thus, by systematically adjusting the parameters of our policy π_θ, the algorithm is capable of effectively compensating for potentially harmful biases by adjusting the severity of the symptoms reported. This leads to an increase in the likelihood of the patient receiving appropriate treatment, as reflected by the policy's convergence towards choosing action a_3 (exaggerating symptoms) when the clinician's initial treatment decision is inadequate.

The positive skew in the reward structure encourages active intervention by the algorithm to correct bias rather than passive non-interference, ensuring that the patient's best interests are prioritized. While this proof establishes the theoretical viability and effectiveness of using a policy gradient method to mitigate clinician bias, it is crucial to consider the overall perspective of such implementations.

6.2 Challenges of Deceptive Dynamics in AI Ecosystems

If too many agents choose deceptive strategies to communicate their preferences, a distorted multi-agent system emerges. Deploying a benevolent deceptive AI in such an environment introduces a multitude of challenges that stem from the complex dynamics of interaction among intelligent entities. These challenges include differing interpretations of fairness, divergent aggressive strategies, inherently malevolent designs, and the potential for misunderstandings or erroneous beliefs regarding other entities' capabilities, intentions, and strategic options. The intricacies of these systems make it overly optimistic to assume that AI entities will autonomously learn to navigate and neutralize deceptive tactics for mutual benefit. Advanced learning capabilities, while impressive, do not inherently equip AI with the means to detect deception or develop counter-deception strategies to support cooperation [88]. As sophisticated agents are prone to falling into traps set by deceptive strategies, this highlights a critical vulnerability reminiscent of human susceptibilities to misinformation and strategic errors. In human interactions, misunderstandings and misinterpretations can lead to conflict; similarly, in AI systems, the inability to

discern deception can undermine trust and cooperation, leading to failures in achieving collective goals.

To address these vulnerabilities, the development of advanced algorithms that incorporate opponent-aware learning strategies becomes essential. These algorithms must explicitly account for the evolving strategies of other agents within the environment, leveraging this understanding to enhance the AI's ability to navigate and counteract deceptive tactics. Furthermore, enhancing communication capabilities between AI systems is crucial for establishing a shared understanding of the strategic environment, which is key to avoiding misinterpretations that could lead to security breaches or failures in cooperation. Effective communication becomes particularly challenging—and vital—in scenarios where adversarial intentions are hidden. Thus, the development of sophisticated methods to ensure clarity and transparency in negotiations, even under potentially adversarial conditions, is imperative.

6.2.1 Significance of Partial Cooperation

Given the numerous challenges in achieving perfect cooperation within any AI ecosystem, especially one involving deceptive agents, aiming for partial cooperation is a more practical objective. This approach is especially pertinent when interacting with agents possessing divergent ethical frameworks or strategic objectives, where full alignment may not be feasible. In these instances, AI systems must be adept at employing a nuanced application of incentives and deterrents to foster a degree of cooperation, ensuring that responses to deception are measured and aim to maintain a baseline of constructive engagement without resorting to or triggering extreme punitive measures, i.e., grim trigger measures that could foreclose the possibility of future cooperation.

6.3 Anticipating User Reactions to Compensatory Adjustments

Now that we are considering the implementation of these systems in the real world, it is important that we also understand how the user might react to compensatory adjustments and what potential long-term solutions might be so as to anticipate challenges, address concerns proactively, and design a system that effectively meets their needs. This understanding will help add nuance to the compensatory adjustments to align with user expectations and foster trust.

6.3.1 Long-Term Effects

Considering the intention behind the compensation is to make the algorithm user a better decision-maker[1], it is essential to study and understand its impact of it on the user. Since the effect would vary from case to case we utilize a flowchart that helps identify the various consequences and explore the dynamics of long-term interactions shown in the figure 7.1.

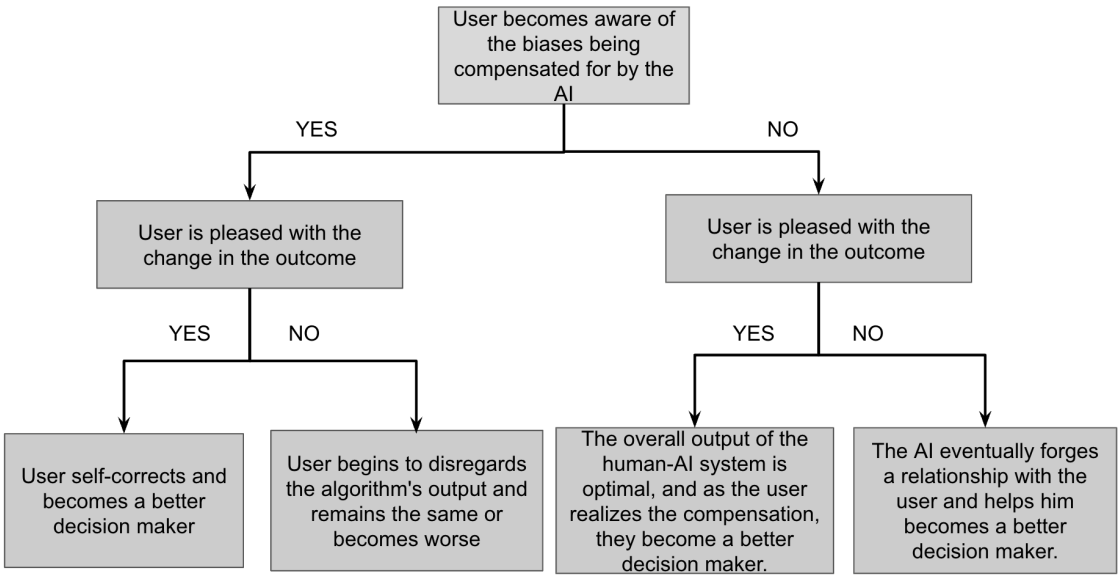

Figure 6.1. Long-term interaction flowchart

[1]refer to introduction section where intervention, inference, and the proposed compensation techniques are mentioned

Intended Consequence: The AI systems are designed to achieve optimal outcomes and improve the decision-maker's (user's) skills and abilities over time. The goal is for the user to learn from the algorithm's suggestions and insights, thereby evolving into a more capable decision-maker. This process ideally creates a positive feedback loop, where the user continually refines their strategies and choices. This scenario often arises when the decision-maker possesses biases that they are unaware of but genuinely wants the best outcomes for those affected by the human-AI system. If the user is open to learning and growing as a decision-maker, the AI can often operate transparently, without resorting to deceptive methods. In these cases, transparency alone is sufficient for the system's success. However, in rare instances where the user's biases influence their decisions, the AI may feel the need to compensate for these occasional lapses to ensure optimal outcomes.

User Dissatisfaction: Deception and Resistance

When designing compensatory systems that interact with humans over prolonged periods, one of the primary considerations must be the human capacity for adaptive learning. Users invariably devise strategies to circumvent or manipulate these systems as they become more familiar with their workings. For instance, in response to changes in social media algorithms that prioritize certain types of posts, users have adapted by using specific hashtags to boost their visibility [89]. This example illustrates the necessity of considering how users might resist or evolve to resist these system. Therefore, when building these systems, it is essential to factor in the potential for such adaptive behaviors to ensure the systems remain effective and relevant over time.

1. Untraced Deception

If the user cannot directly attribute their dissatisfaction with outcomes to the AI system's deception, they might resort to suboptimal behaviors more intensely instead of reconsidering their approach. In such cases, the user's actions could continue deteriorating as they act on

misleading information or flawed assumptions.

Algorithm Response: The algorithm might respond by increasing the level of deception, hoping to balance the user's decision-making behavior. This escalation aims to redirect the user toward better decisions without revealing the deceptive strategy at play. However, this escalation can lead to a 'Deception Loop,' where:

- The user's actions continue to worsen, adhering stubbornly to poor decisions.
- The algorithm increases its level of deception to balance the user's behavior.

This loop is unsustainable because, at a certain point, deception cannot be increased without risking detection. To restore trust and cooperation, the algorithm must carefully reduce its deceptive measures to subtly nudge users toward optimal behavior while maintaining credibility.

2. User Discovers Source of Deception

If the user traces these manipulative actions back to the system, they may view it as deceptive. This realization can lead to mistrust and non-compliance, prompting the user to disregard the algorithm's guidance entirely. Much like the Grim Trigger strategy, where a single act of betrayal prompts ongoing retaliation, the user may decide to exclude the AI's input from their decision-making process as a form of punitive response. This is also in line with popular research in algorithm aversion [90, 91] where "people often fail to use (algorithms) after learning that they are imperfect."

To prevent this, the AI must prioritize maintaining a trustworthy relationship by ensuring that any deceptive elements remain subtle and cannot be traced back to the system.

6.3.2 Long-Term Solutions

Standard measures like education to raise awareness, ensuring diversity and transparency at every level of an organization, and establishing effective grievance resolution processes can

gradually help address underlying issues. However, these alone are insufficient, as their impact takes time to manifest and may not be as effective as intended.

Complementing these measures with a compensation algorithm could prove more impactful. Such an algorithm would continuously identify and address biases as they emerge, acting as a permanent fixture that promotes rational, objective decision-making.

6.4 Acknowledgement

This chapter, in part, is a reprint of the material presented at the International Association for Computing and Philosophy, 2022 (Swaminathan, N., & Danks, D. (2023). When Can My AI Lie? (No. 10063). EasyChair.). It has also has been submitted for publication and is currently under review. The thesis author was the primary investigator and author of this paper.

Chapter 7

Conclusion

We have argued that as artificial learning agents become increasingly widespread in our society, there is a growing need to navigate complex interactions with other agents (both human and nonhuman) in the environment. We have also demonstrated that these interactions may sometimes involve elements of deception, raising important ethical questions. To address this challenge, we propose that further research at the intersection of game theory and artificial intelligence is necessary to develop methods and techniques that allow AI systems to utilize deception in an ethical manner to navigate social dilemmas productively.

In this work, we have considered the perspective of an AI agent within a human-AI system, focusing on how the agent can promote overall system performance by compensating for human biases and ensuring fair outcomes for the target of the decisions. We have proposed a framework outlining how the agent could learn to provide these compensatory adjustments by considering the target's anticipated consent and the user's patterns of bias.

However, our proposed framework is just a starting point, and there is much room for further development and refinement. Future research on compensatory bias-reducing mechanism design could explore several key areas:

- Developing more sophisticated models for inferring and representing the target's preferences and consent.
- Conducting empirical studies and simulations to test the effectiveness and robustness of

different compensatory bias-reducing mechanisms in real-world settings.

In this work, we also argue that the overarching goal of multi-agent learning should be the development of AI systems that are designed to work effectively with other agents (both human and artificial) towards the realization of mutually beneficial outcomes. The framework and insights presented in this thesis, while preliminary, aim to contribute to the growing field of research on AI. By proposing a mechanism for AI agents to learn to compensate for human biases and promote fair outcomes through the judicious use of deception, we hope to stimulate further research and discussion on the complex ethical challenges that arise in human-AI collaboration.

However, we recognize that this work represents only a modest step towards the ambitious goal of building truly cooperative AI systems. Much further research is needed to develop sophisticated models of multi-agent interaction that can be deployed in real-world contexts.

Bibliography

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, and S. . a. a. n. s. Anadkat, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] A. Azaria and T. Mitchell, "The internal state of an llm knows when its lying," *arXiv preprint arXiv:2304.13734*, 2023.

[4] J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, P. J. Bentley, S. Bernard, G. Beslon, D. M. Bryson, *et al.*, "The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities," *Artificial life*, vol. 26, no. 2, pp. 274–306, 2020.

[5] J. E. Korteling, G. L. Paradies, and J. P. Sassen-van Meer, "Cognitive bias and how to improve sustainable decision making," *Frontiers in Psychology*, vol. 14, p. 1129835, 2023.

[6] M. Bertrand and S. Mullainathan, "Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination," *American economic review*, vol. 94, no. 4, pp. 991–1013, 2004.

[7] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, vol. 133, no. 1, pp. 237–293, 2018.

[8] C. Wang, K. Wang, A. Y. Bian, R. Islam, K. N. Keya, J. Foulds, and S. Pan, "When biased humans meet debiased ai: A case study in college major recommendation," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 3, pp. 1–28, 2023.

[9] Y.-T. Lin, T.-W. Hung, and L. T.-L. Huang, "Engineering equity: How ai can help reduce the harm of implicit bias," *Philosophy & Technology*, vol. 34, no. Suppl 1, pp. 65–90, 2021.

[10] E. Drage and K. Mackereth, "Does ai debias recruitment? race, gender, and ai's "eradication of difference"," *Philosophy & technology*, vol. 35, no. 4, p. 89, 2022.

[11] A. Rizer and C. Watney, "Artificial intelligence can make our jail system more efficient, equitable, and just," *Tex. Rev. L. & Pol.*, vol. 23, p. 181, 2018.

[12] D. Wagner, "On the emergence and design of ai nudging: the gentle big brother?," *ROBONOMICS: The Journal of the Automated Economy*, vol. 2, pp. 18–18, 2021.

[13] R. H. Brescia, "On tipping points and nudges: Review of cass sunstein's how change happens," *Notre Dame JL Ethics & Pub. Pol'y*, vol. 34, p. 55, 2020.

[14] S. Cremin, "Nudging judges away from implicit bias: Using behavioral science to promote racial equity in federal sentencing," *New Eng. L. Rev.*, vol. 56, p. 57, 2021.

[15] C. R. Sunstein, "Nudges that fail," *Behavioural public policy*, vol. 1, no. 1, pp. 4–25, 2017.

[16] P. Masters, W. Smith, L. Sonenberg, and M. Kirley, "Characterising deception in ai: A survey," in *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*, pp. 3–16, Springer, 2021.

[17] J. Schneider, C. Meske, and M. Vlachos, "Deceptive ai explanations: Creation and detection," *arXiv preprint arXiv:2001.07641*, 2020.

[18] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, "Ai deception: A survey of examples, risks, and potential solutions," *arXiv preprint arXiv:2308.14752*, 2023.

[19] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pp. 101–108, IEEE, 2016.

[20] E. E. Levine and M. E. Schweitzer, "Are liars ethical? on the tension between benevolence and honesty," *Journal of Experimental Social Psychology*, vol. 53, pp. 107–117, 2014.

[21] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social justice research*, vol. 20, no. 1, pp. 98–116, 2007.

[22] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[23] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, pp. 1889–1897, PMLR, 2015.

[24] S. Cohen, "Manipulation and deception," *Australasian Journal of Philosophy*, vol. 96, no. 3, pp. 483–497, 2018.

[25] T. L. Carson, *Lying and deception: Theory and practice*. OUP Oxford, 2010.

[26] L.-M. Russow, "Deception: A philosophical perspective," *Deception: Perspective on human and nonhuman deceit*, pp. 41–52, 1986.

[27] J. E. Mahon, "The definition of lying and deception," 2008.

[28] S. Bodmer, D. M. Kilger, G. Carpenter, J. Jones, and J. Jones, *Reverse deception: organized cyber threat counter-exploitation*. McGraw-Hill New York, 2012.

[29] H. B. Cott, *Adaptive coloration in animals*. London: Methuen, 1940. University of Florida, George A. Smathers Libraries.

[30] U. Gneezy, "Deception: The role of consequences," *American Economic Review*, vol. 95, no. 1, pp. 384–394, 2005.

[31] A. Vrij, S. A. Mann, R. P. Fisher, S. Leal, R. Milne, and R. Bull, "Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order," *Law and human behavior*, vol. 32, pp. 253–265, 2008.

[32] F. Cuzzolin, A. Morelli, B. Cirstea, and B. J. Sahakian, "Knowing me, knowing you: theory of mind in ai," *Psychological medicine*, vol. 50, no. 7, pp. 1057–1061, 2020.

[33] P. Felli, T. Miller, C. Muise, A. R. Pearce, and L. Sonenberg, "Artificial social reasoning: computational mechanisms for reasoning about others," in *Social Robotics: 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014. Proceedings 6*, pp. 146–155, Springer, 2014.

[34] C. Langley, B. I. Cirstea, F. Cuzzolin, and B. J. Sahakian, "Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review," *Frontiers in Artificial Intelligence*, vol. 5, p. 62, 2022.

[35] D. Premack, "Premack and woodruff: Chimpanzee theory of mind," *Behavioral and Brain Sciences*, vol. 4, no. 1978, pp. 515–526, 1978.

[36] D. C. Dennett, "Intentional systems," *The Journal of Philosophy*, vol. 68, no. 4, pp. 87–106, 1971.

[37] N. J. Block, "Functionalism," in *Studies in Logic and the Foundations of Mathematics*, vol. 104, pp. 519–539, Elsevier, 1982.

[38] S. Shoemaker, "Some varieties of functionalism," *Philosophical topics*, vol. 12, no. 1, pp. 93–119, 1981.

[39] B. A. Levinstein and D. A. Herrmann, "Still no lie detector for language models: Probing empirical and conceptual roadblocks," *Philosophical Studies*, pp. 1–27, 2024.

[40] A. Alabed, A. Javornik, and D. Gregory-Smith, "Ai anthropomorphism and its effect on users' self-congruence and self–ai integration: A theoretical framework and research agenda," *Technological Forecasting and Social Change*, vol. 182, p. 121786, 2022.

[41] M. L. Littman, "Reinforcement learning improves behaviour from evaluative feedback," *Nature*, vol. 521, no. 7553, pp. 445–451, 2015.

[42] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[43] K. Tuyls and G. Weiss, "Multiagent learning: Basics, challenges, and prospects," *Ai Magazine*, vol. 33, no. 3, pp. 41–41, 2012.

[44] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 750–797, 2019.

[45] S. Sen, M. Sekaran, and J. Hale, "Learning to coordinate without sharing information," in *AAAI*, vol. 94, pp. 426–431, 1994.

[46] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*, pp. 157–163, Elsevier, 1994.

[47] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," *AAAI/IAAI*, vol. 1998, no. 746-752, p. 2, 1998.

[48] D. Floreano, S. Mitri, S. Magnenat, and L. Keller, "Evolutionary conditions for the emergence of communication in robots," *Current biology*, vol. 17, no. 6, pp. 514–519, 2007.

[49] S. Mitri, D. Floreano, and L. Keller, "Evolutionary conditions for the emergence of communication," *Evolution of Communication and Language in Embodied Agents*, pp. 123–134, 2010.

[50] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra, "Deal or no deal? end-to-end learning for negotiation dialogues," *arXiv preprint arXiv:1706.05125*, 2017.

[51] M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, *et al.*, "Human-level play in the game of diplomacy by combining language models with strategic reasoning," *Science*, vol. 378, no. 6624, pp. 1067–1074, 2022.

[52] O. Weisel and S. Shalvi, "The collaborative roots of corruption," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. 10651–10656, 2015.

[53] S. S. Wiltermuth, L. C. Vincent, and F. Gino, "Creativity in unethical behavior attenuates condemnation and breeds social contagion when transgressions seem to create little harm," *Organizational Behavior and Human Decision Processes*, vol. 139, pp. 106–126, 2017.

[54] Y. Mao, B. Hu, and K. J. Kim, "When ai doctors lie about diagnosis: The effects of varying degrees of prosocial lies in patient–ai interactions," *Technology in Society*, vol. 76, p. 102461, 2024.

[55] T. Chakraborti and S. Kambhampati, "(when) can ai bots lie?," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 53–59, 2019.

[56] S. Sarkadi, P. Mei, and E. Awad, "Should my agent lie for me? a study on attitudes of us-basedparticipants towards deceptive ai in selected future-of-work," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 345–354, 2023.

[57] J. R. Auman, *Cyber war tactics or clever behavior: Understanding cyber deception techniques in the fight against cyber warfare*. PhD thesis, Utica College, 2014.

[58] K. D. Martinez, *Challenge of applying tactical deception when conducting large scale combat operations in the 21st Century*. PhD thesis, Fort Leavenworth, KS: US Army Command and General Staff College, 2021.

[59] E. Chelioudakis, "Deceptive ai machines on the battlefield: Do they challenge the rules of the law of armed conflict on military deception?," *Available at SSRN 3158711*, 2017.

[60] M. Lloyd, *The art of military deception*. Pen and Sword, 2003.

[61] A. R. Wagner and R. C. Arkin, "Acting deceptively: Providing robots with the capacity for deception," *International Journal of Social Robotics*, vol. 3, pp. 5–26, 2011.

[62] A. Isaac and W. Bridewell, "Why robots need to deceive (and how)," *Robot ethics*, vol. 2, pp. 157–172, 2017.

[63] W. Bridewell and A. Isaac, "Recognizing deception: A model of dynamic belief attribution," in *2011 AAAI Fall Symposium Series*, 2011.

[64] K. Vaccaro, D. Huang, M. Eslami, C. Sandvig, K. Hamilton, and K. Karahalios, "The illusion of control: Placebo effects of control settings," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.

[65] J. Shim and R. C. Arkin, "Biologically-inspired deceptive behavior for a robot," in *International conference on simulation of adaptive behavior*, pp. 401–411, Springer, 2012.

[66] E. de Oliveira, L. Donadoni, S. Boriero, and A. Bonarini, "Deceptive actions to improve the attribution of rationality to playing robotic agents," *International Journal of Social Robotics*, vol. 13, pp. 391–405, 2021.

[67] B. R. Brewer, R. L. Klatzky, and Y. Matsuoka, "Visual-feedback distortion in a robotic rehabilitation environment," *Proceedings of the IEEE*, vol. 94, no. 9, pp. 1739–1751, 2006.

[68] W. M. C. M. M. A. D. J. C. D. H. C. R. P. T. E. P. G. J. O. D. H. M. K. I. D. A. H. L. S. T. C. J. P. A. M. J. A. S. V. R. L. T. P. V. D. D. B. Y. S. J. M. T. L. P. C. G. Z. W. T. P. Y. W. R. R. D. Y. D. W. K. M. O. S. T. S. T. L. K. K. D. H. C. A. . D. S. Oriol Vinyals, Igor Babuschkin, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[69] K. S. I. A. A. H. A. G. T. H. L. B. M. L. A. B. Y. C. T. L. F. H. L. S. G. v. d. D. T. G. D. H. David Silver*, Julian Schrittwieser*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[70] D. S. A. A. R. J. V. M. G. B. A. G. M. R. A. K. F. G. O. S. P. C. B. A. S. I. A. H. K. D. K. D. W. S. L. . D. H. Volodymyr Mnih, Koray Kavukcuoglu, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[71] B. Sjödén, "When lying, hiding and deceiving promotes learning-a case for augmented intelligence with augmented ethics," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pp. 291–295, Springer, 2020.

[72] X. Zhai, X. Chu, C. S. Chai, M. S. Y. Jong, A. Istenic, M. Spector, J.-B. Liu, J. Yuan, and Y. Li, "A review of artificial intelligence (ai) in education from 2010 to 2020," *Complexity*, vol. 2021, pp. 1–18, 2021.

[73] F. Tanaka and T. Kimura, "Care-receiving robot as a tool of teachers in child education," *Interaction Studies*, vol. 11, no. 2, p. 263, 2010.

[74] S. Matsuzoe and F. Tanaka, "How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pp. 339–344, IEEE, 2012.

[75] A. Furnham and H. C. Boo, "A literature review of the anchoring effect," *The journal of socio-economics*, vol. 40, no. 1, pp. 35–42, 2011.

[76] M. J. Osborne, *An introduction to game theory*, vol. 3. Oxford university press New York, 2004.

[77] S. Darwall, "The value of autonomy and autonomy of the will," *Ethics*, vol. 116, no. 2, pp. 263–284, 2006.

[78] J. Kleinig, *Paternalism*. Manchester University Press, 1983.

[79] C. H. Schroeder, "Rights against risks," *Colum. L. Rev.*, vol. 86, p. 495, 1986.

[80] B. McElwee, "The rights and wrongs of consequentialism," *Philosophical Studies*, vol. 151, pp. 393–412, 2010.

[81] A. Nelson, "Unequal treatment: confronting racial and ethnic disparities in health care.," *Journal of the national medical association*, vol. 94, no. 8, p. 666, 2002.

[82] D. R. Williams, "Miles to go before we sleep: Racial inequities in health," *Journal of health and social behavior*, vol. 53, no. 3, pp. 279–295, 2012.

[83] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ digital medicine*, vol. 3, no. 1, p. 17, 2020.

[84] S. Holm, "Principles of biomedical ethics," *Journal of Medical Ethics*, vol. 28, no. 5, p. 332, 2002.

[85] D. K. Sokol, "Can deceiving patients be morally acceptable?," *BMJ*, vol. 334, no. 7601, pp. 984–986, 2007.

[86] J. Hoberman, "Medical racism and the rhetoric of exculpation: how do physicians think about race?," *New Literary History*, vol. 38, no. 3, pp. 505–525, 2007.

[87] T. L. Beauchamp and J. F. Childress, *Principles of biomedical ethics*. Oxford University Press, USA, 2001.

[88] Ş. Sarkadi, A. Rutherford, P. McBurney, S. Parsons, and I. Rahwan, "The evolution of deception," *Royal Society open science*, vol. 8, no. 9, p. 201032, 2021.

[89] M. Eslami, K. Vaccaro, K. Karahalios, and K. Hamilton, ""be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, pp. 62–71, 2017.

[90] E. Jussupow, I. Benbasat, and A. Heinzl, "Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion," 2020.

[91] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management science*, vol. 64, no. 3, pp. 1155–1170, 2018.