Behavioral economics for AI-augmented scenarios

Tapio Tuloisela

School of Science

Bachelor's thesis
Espoo 16.9.2024

Supervisor

Prof. Lauri Savioja

Advisor

Asst. Prof. Robin Welsch

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Tapio Tuloisela

Title Behavioral economics for AI-augmented scenarios

Degree programme Bachelor's Programme in Science and Technology

Major Engineering Psychology **Code of major** SCI3163

Teacher in charge Prof. Lauri Savioja

Advisor Asst. Prof. Robin Welsch

Date 16.9.2024 **Number of pages** 33 **Language** English

Abstract

The applications of Artificial Intelligence (AI) are becoming increasingly prominent in the society. Therefore, it is valuable to develop a broader understanding of human decision-making in AI-augmented scenarios. For instance, independent AI agents and other advanced AI applications, such as digital assistants are becoming ubiquitous. A deep understanding of human-AI interactions is important in order to ensure that the development of AI-driven society takes into account the possible challenges that might arise from this new field of interaction.

This bachelor's thesis is a literature review exploring how humans cooperate and how these cooperative traits transition to human-AI strategic interactions. The concepts are studied using the methods of behavioral economics and game theory that have shown good results in predicting human decision-making in strategic interactions. The goal of the thesis is to inspect how human behavior changes in AI-augmented scenarios.

Real-life strategic interactions can be represented with game theoretic scenarios called economic games. A common variation of these games are social dilemmas that enable researchers to study various motivations behind economic decision-making. In empirical studies on economic games, mathematical models that take into account people's preferences for fairness and cooperation tend to outperform predictions focused solely on self-interested motivations. Typically these models compare decisions by assigning utilities for different actions based on individual's social preferences and material rewards in the scenario. Examining these models is useful, because they provide insight into humans' internal decision-making processes.

The thesis concludes that in AI-augmented strategic interactions, the current social preference models are unable to predict cooperation accurately. Instead, people seem to exploit benevolent AI agents. However, social preferences tend to persist when individuals are informed that there is another human beneficiary who receives the payoffs gained by the AI. This also seems to translate to settings where humans are interacting with other AI-augmented people. Even with no human beneficiary, cooperation can be elicited by introducing cultural or emotional expressions for the AI agents. Based on the reviewed literature, the thesis presents a framework to evaluate human decision-making in strategic interactions, which also considers scenarios involving non-human agents.

Keywords artificial intelligence, behavioral economics, game theory, strategic interaction

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Tapio Tuloisela

Työn nimi Käyttäytymistaloustiede tekoälyä sisältävissä vuorovaikutustilanteissa

Koulutusohjelma Teknistietellinen kandidaattiohjelma

Pääaine Teknillinen psykologia **Pääaineen koodi** SCI3163

Vastuuopettaja Prof. Lauri Savioja

Työn ohjaaja Asst. Prof. Robin Welsch

Päivämäärä 16.9.2024 **Sivumäärä** 33 **Kieli** Englanti

Tiivistelmä

Tekoälyn sovellutukset ovat integroitumassa yhä laajemmin osaksi yhteiskuntaa. Tämän seurauksena osoittautuu tärkeäksi ymmärtää ihmisten käyttäytymistä tekoälyn kanssa vuorovaikutettaessa. Esimerkiksi itseajavat autot ja edistyneet generatiivisen tekoälyn sovellukset, kuten keskustelubotit valtaavat yleistä kuvaa tulevaisuudesta. Syvällisellä ymmärryksellä ihmisen ja tekoälyn vuorovaikutuksesta voidaan parantaa yhteiskunnan turvallisuutta ja soveltaa tekoälyn kehitystä ihmisystävällisemmän tulevaisuuden takaamiseksi.

Tämä kandidaatintyö käsittelee kirjallisuuskatsauksen kautta ihmisten välisen yhteistyön juurisyitä strategisessa vuorovaikutuksessa sekä tekoälyelementtien tuomaa muutosta yhteistyön esiintymiseen. Työssä tutkitaan erityisesti behavioristisen peliteorian menetelmiä, sillä aiempi tutkimus käyttäytymistaloustieteen alalla on osoittanut niiden olevan erinomainen tapa luoda ennusteita ihmisten päätöksenteolle monimutkaisissakin tilanteissa. Työn tavoite on tarkastella, miten asenteet tekoälytoimijoita tai muita tekoälytyökaluja käyttäviä ihmisiä kohtaan vaikuttavat yhteistyön esiintymiseen strategisissa vuorovaikutustilanteissa.

Ihmisten välisen yhteistyön mallintamiseen on kehitetty peliteoreettisia skenaarioita, jotka mallintavat yksinkertaistetusti toimijoiden välisiä tosielämän strategisia vuorovaikutustilanteita. Tyypillisiä asetelmia ovat ns. sosiaaliset dilemmat, joiden avulla voidaan tutkia erilaisia valintojen taustalla vaikuttavia motiiveja. Vaihtelevat tilanteet tarjoavat tietoa ihmisten sisäisistä motiiveista eri tavoin. Esimerkiksi niin kutsutussa metsästyspelissä (engl. Stag hunt) osapuolten päätöksentekoon vaikuttavat kateus tai epäluottamus pelaajien välillä. Sen sijaan luottamuspeleissä (engl. Trust games) jälkimmäisen pelaajan päätös yhteistyön tekemisestä kertoo hänen motiiveistaan yhteisen hyvän ja itsekkyyden painottamisen välillä. Empiirisissä tutkimuksissa matemaattiset mallit, jotka ottavat huomioon yksilöiden sosiaaliset mieltymykset esimerkiksi epätasa-arvon välttämisen suhteen ovat osoittautuneet ennustekyvyltään tehokkaammiksi kuin puhtaasti aineellisia palkintoja huomioivat ennusteet. Mallien tarkastelu on hyödyllistä, sillä ne tarjoavat tietoa ihmisen sisäisistä päätöksentekoprosesseista.

Työssä todetaan, että sosiaalisten mieltymysten mallit eivät nykyisellään kykene ennustamaan yhteistyön esiintymistä tekoälyätoimijoiden kanssa vuorovaikutettaessa. Sen sijaan ihmisten havaitaan usein hyväksikäyttävän yhteistyökykyisiä tekoälytoimijoita. Toisaalta yhteistyöhalukkuus vaikuttaa kuitenkin riippuvan myös tekoälyn implementointiin liittyvistä seikoista. Perinteisten yhteistyöhön vaikuttavien ele-

menttien lisäksi tekoälyn kanssa vuorovaikutettaessa esimerkiksi ihmisedunsaajan implikointi tai tunneilmaisuja välittävän avatarin käyttö tekoälytoimijan yhteydessä todetaan lisäävän sosiaalisia mieltymyksiä. Alan tutkimustietoon perustuen työssä esitellään viitekehys, jonka avulla yhteistyön esiintymistä voidaan arvioida niin ihmisten kuin tekoälytoimijoidenkin kanssa vuorovaikutettaessa.

Avainsanat tekoäly, käyttäytymistaloustiede, peliteoria, strateginen vuorovaikutus

Contents

1 Introduction

As the role of artificial intelligence (AI) systems is becoming more important in the society, new questions arise about the impact of AI presence on human behavior. For instance in traffic, human attitudes toward self-driving cars may differ from conventional interactions with other human-drivers. In recent years, artificial intelligence and its applications have become a popular concept of debate. In spite of this, AI has been an important part of the society for quite some time. In one of the earliest publications related to AI, McCulloch and Pitts (1943) proposed a computer model that can learn in a similar fashion as human neurons. Since then, the development of AI has seen cycles of high expectations and eventual setbacks (Muthukrishnan et al., 2020). In 2010s and 2020s the progress has been unprecedented (Aschenbrenner, 2024). AI is being integrated to many parts of people's everyday lives: digital assistants, self-parking and self-driving cars, predictive entries in search engines, chatbots and more subtly, algorithms to target advertisements or recommend the most attractive content for the user in social media (Stone et al., 2016). AI-augmentation offers help with analytical tasks: large language models seem to be able to evaluate human emotional states and aid mental health assessment by combining data from multiple sources Hu et al. (2024).

The development of AI systems brings new problems as well. For example, AI models might inherit bias present in their training data, thus continuing to reinforce the bias as people interact with the AI (Hacker, 2018). There has been critique towards AI developers for not taking societal concepts such as social bias into account, however this has recently been changing. Social scientists are now starting to be considered an important source of expertise by AI researchers (Kusner and Loftus, 2020).

In a world where AI systems are becoming more prominent, it is important to examine how people behave in situations involving AI agents or augmentation compared to regular interactions with other human beings. Past research in economic games suggests that people often cooperate with others even when acting in self-interest would not damage their reputation (Johnson and Mislin, 2011). However, in the case of AI agents, people tend to exploit benevolent AI for selfish gains (Karpus et al., 2021). What this would potentially mean to how humans cooperate with self-driving cars in traffic for example is an important question.

Game theory used to be a branch of mathematics studying how rational players can achieve the highest material gain interacting with other players. In recent decades the field has evolved to become applicable also to mundane, often irrational human encounters. Behavioral game theory provides accurate means for researching strategic interaction and predicting how humans make decisions (Camerer, 1997). For instance, behavioral game theory may be utilized to find the optimal strategy to gain influence in a variety of social settings (Kurvers et al., 2021). This suggests that when applied correctly, game theory can describe even more complex scenarios that social interactions represent. Therefore, behavioral game theory is a useful tool for deeper understanding of human-AI interaction.

The purpose of this thesis is to study how humans cooperate and how the cooperative traits transition to human-AI strategic interactions. The thesis is organised as follows. Section 2 explains the relevant background for the thesis about behavioral economics and human-AI interaction. Section 3 presents the game theoretic concepts for human cooperation rooted in evolution and how these traits manifest today as social preferences. Section 4 reviews literature on human-AI strategic interactions and presents a framework for how humans form cooperative decisions. Finally, section 5 concludes the thesis. One may find the thesis useful when studying explanations for varying behavioral or emotional reactions in human and human-AI interaction scenarios.

2 Background

This section covers important concepts and definitions for the thesis. The section also presents related work on the topics.

2.1 Behavioral economics

Mullainathan and Thaler (2000) define behavioral economics as "the combination of psychology and economics that investigates what happens in markets in which some of the agents display human limitations and complications." Humans tend to attribute subjective value for economic decisions beyond pure rational interest. For instance, people give more subjective value for losing than gaining an equal amount of assets

(Kahneman and Tversky, 1979) and dislike inequally distributed resources within their social groups (Fehr and Schmidt, 1999). Essentially, in contrast to limiting the study of economics to rational agents, the role of behavioral economics is to introduce irrational human tendencies to the equation of economic decision-making.

Behavioral economics and its game theoretic applications are not solely interested in scenarios involving money or trading of assets. In economic literature, the concept of value is often based on personal evaluation where for example functional and social aspects are also considered (Boksberger and Melsen, 2011). Therefore any situation where an individual is acquiring subjective value while interacting with other agents, i. e. strategic interactions, are of interest. Such acquisition of value might as well be higher genetic fitness or reduced waiting time in traffic.

Game theory is a field of applied math used to study how interacting decision-makers with varying motives behave in strategic settings (Osborne, 1994). In traditional game theory, the agents are often assumed to be self-interested and to possess information of other decision-maker's actions and motives. Moreover, Nash equilibrium is a central concept of game theory which is defined as a state where no agent can get more preferable outcome by changing their action (Osborne, 1994, p.14). Extending the field to be more accurate for real-life interactions, the role of behavioral game theory is to provide mathematical explanations to how people act in scenarios where the empirical evidence contradicts self-interest model's predictions (Camerer, 1997).

Suppose a game of ultimatum, a game where one player proposes a share of a reward between themselves and another player. The second player can then choose if they accept the division or reject it resulting in both players getting zero reward. For ultimatum game, the traditional game theory predicts that a self-interested responder will be indifferent to accept or reject the offer if the first player proposes to keep all of the reward for themselves (Nowak et al., 2000). If the first player decides to give even an arbitrarily small amount of the reward to the second player, the division will be accepted. The second player is expected to be happy because instead of getting nothing, they receive a non-zero amount. Empirical evidence shows that the prediction of self-interest model does not hold and people playing as the second player tend to reject offers they perceive as unfair (Nowak et al., 2000). Thus, behavioral game theory brings an important aspect of varying motivations into the equation of strategic interactions.

2.2 Human-AI interaction

Artificial intelligence refers to algorithm-based computer applications designed to approximate intelligence and make predictions based on its training data (Alkatheiri, 2022). AI can generally be divided into two categories: machine-learning and generative AI. Especially the development of generative AI has been rapid in the past. Between 2019 and 2023 the content produced by large language models advanced from the level of a child to a smart high-school student and the predictions to reach an expert level vary from 4-10 years (Aschenbrenner, 2024). The definition for AI-augmentation varies among disciplines but it is often referred to as humans using AI tools to enhance their performance (Sadiku and Musa, 2021). In this thesis, AI-augmented strategic interaction refers to a situation of economic interests which in addition to humans, involves AI agents or other AI-augmented humans.

It is apparent that several tasks can be made more efficient using AI enhancement. For instance, professionals with GPT-3.5 assistance were shown to be more productive and have improved quality in work-related tasks (Noy and Zhang, 2023). However, as people are not traditionally used to interact with non-human entities, the AI presence may also result in unexpected problems. Means to study the possible challenges caused by the generalization of AI systems exist. Social dilemma games with manipulated action space offer information about underlying motivations in strategic interactions and can be used to study how people interact with AI systems (Karpus et al., 2021; von Schenk et al., 2023; de Melo and Terada, 2019).

How the cooperative tendencies of humans are altered when they are interacting with another AI-augmented person or AI agents is unsettled scientifically. Research suggests that when playing sequential social dilemma games, expert algorithms can even exceed cooperation rates with humans compared to humans playing with other humans (Crandall et al., 2018). However, when people are aware of the non-human nature of the agent that they are interacting with, cooperation tends to decrease (Ishowo-Oloko et al., 2019). Human attitudes towards AI agents are also connected to how opaque or emotionless the AI system is perceived to be (De Freitas et al., 2023). Thus, experiments that increase relatedness towards machines have shown positive results in making people more cooperative with non-human agents (de Melo and Terada, 2019). Each individual's cultural background may also play a role as for instance eastern cultures tend to perceive AI systems more often as having a spirit or a soul (De Freitas et al., 2023).

3 Game theoretic approach to cooperation

The phenomenon of cooperation has long puzzled researchers. In an environment where exploiting one's partner yields higher benefit, an question raises of how evolution has sustained the cooperative traits that are still natural for most people. This section covers the most important social dilemma games from economic literature that reveal varying motivations behind the agents' actions. Theories associated with the rise of cooperation in evolution are also presented. These cooperative traits manifest today as social preferences even in strategic interactions where cooperation cannot be explained by self-interest.

3.1 Cooperative dilemmas

The problem of cooperation can be examined with economic games that model scenarios of cooperative behavior between individuals. The following 2x2 matrix visualizing the payoffs for a two player scenario depending on decisions A and B has been a workhorse of microeconomics studies for decades.

	C	D
C	R	S
D	T	P

Table 1: Payoff matrix indicating the row player's payoffs where R = reward for mutual cooperation, T = temptation to defect, S = sucker's payoff and P = punishment for mutual defection

When the following conditions hold (Nowak et al., 2012), it makes sense to label the choices as C = cooperation, D = defection: (1.) both players cooperate, they obtain a larger payoff than by mutually defecting $R > P$ but (2.) there is still incentive for defection. This incentive may emerge in three different ways. If $T > R$, the player receives a higher payoff by defecting compared to cooperating when playing against a cooperator. If $P > S$, then it is preferable to defect against a defector and if $T > S$, it is better to be the defector in a scenario featuring a defector and a cooperator.

The advantage of social dilemma games is that they provide a framework to abstractly evaluate economic decision-making and motivations behind different economic actions.

The cooperative and defective decisions in the following games may be motivated differently depending on the game's payoff structure. Therefore studying each game provides different information about human motivations in strategic interactions.

3.1.1 Prisoner's dilemma

When each of the social dilemma conditions are true and $T > R > P > S$, the game is called a prisoner's dilemma.

	C	D
C	$70, 70$	$0, 100$
D	$100, 0$	$30, 30$

Table 2: An example payoff matrix for prisoner's dilemma

As table 2 illustrates, in prisoner's dilemma $T > R$ and $P > S$ resulting in the fact that it is always better for both of the participants to defect regardless of what the other player does. Therefore many of the theories explaining cooperation argue that only for prisoner's dilemmas that are played sequentially, it is in the self-interests of the participants to cooperate. By exploiting cooperation, they induce the opposing player to defect in the following rounds. Thus cooperating and building mutual trust will be preferred even for agents that are not intrinsically altruistic.

3.1.2 Chicken game

	C	D
C	$70, 70$	$20, 100$
D	$100, 20$	$0, 0$

Table 3: Example of a possible payoff structure in the chicken game

Another 2x2 matrix game is the chicken game. This game is identical to Prisoner's dilemma with the exception that $S > P$. For Chicken game, the cooperation rates between humans are usually higher than in prisoner's dilemma because in this case the payoff for mutual defection is the lowest. Just as prisoner's dilemma, chicken game reveals player's greedy preferences as the highest payoff can be achieved by

defecting against a cooperator. However, the cost of possible lowest payoff resulting from mutual defection lures players to cooperate more.

3.1.3 Stag hunt

Stag hunt is a social dilemma with different motivational background from the previously presented games where $R > T$ and $P \geq S$.

	C	D
C	$100, 100$	$0, 50$
D	$50, 0$	$50, 50$

Table 4: An example of stag hunt payoff structure

Table 4 illustrates that it is in both players' interest to cooperate in order to get the highest payoff. The incentive to defect results from suspecting that the opponent might defect, which would yield a low payoff for the cooperator. Thus, a lack of trust between the participants may cause a player to defect and get a guaranteed moderate payoff. Another possible motivation for defection in stag hunt is that the defecting player has a preference to "win" their opponent, i. e. get a higher payoff in the event that their opponent cooperates.

3.1.4 Trust games

Another branch of economic games typically assessed to study human cooperation are called trust games or gift-exchange games. Compared to 2x2 matrix games, trust game are asymmetric. This means that the moves are made sequentially and at least one player makes their choice after having full information whether the first player was cooperative or not. If the first player cooperates, they trust the other player to be cooperative in return.

A possible scenario of trust game is presented in figure 1. Player 1 decides whether to "trust" the second player and give up their guaranteed medium payoff and cooperate. Player 2 may then choose to cooperate and choose an option where both players get a higher reward. They also have a possibility to choose the temptative larger personal reward and exploit the first player's trust.

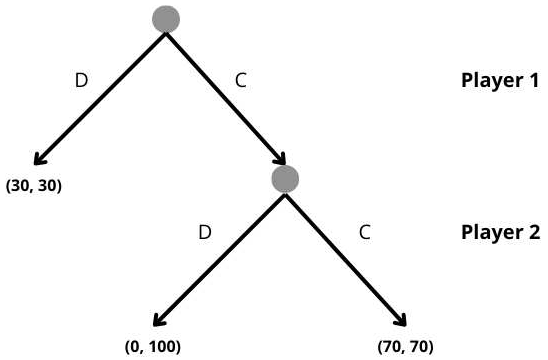


Figure 1: An example of trust game with binary decision space

The traditional game theoretic self-interest model predicts that second players will always choose the exploitative option. Therefore game theory also states that the first player should never cooperate. Experimental evidence with human players suggests that the traditional game theory model is not sufficient to predict how people behave in this scenario. In experiments between humans, many player 2 participants tend to cooperate and give up their larger payoff to reciprocate the trust from player 1 (Karpus et al., 2021).

3.1.5 Overview of the games

Game	Symmetric	Motivations
Prisoner's Dilemma	Yes	Fear of exploitation. Temptation to maximize personal gain.
Chicken Game	Yes	Fear of mutual defection. Temptation to maximize personal gain
Stag Hunt	Yes	Trust/distrust in opponent's cooperation. Willingness to outperform opponent.
Trust Game 1st player	No	Trust/distrust that the second player will reciprocate cooperation.
Trust Games 2nd player	No	Self-interest to maximize gain. Guilt of exploiting 1st player's trust.

Table 5: Comparison of economic games

Table 5 illustrates the decision spaces of each game and the underlying motivations behind the actions. As shown, each game models different motivations and can be used to gather varying information about the decision-making processes of the players. By combining data from multiple social dilemma games, these scenarios can be used to construct a deeper understanding of attitudes and bias between groups.

3.2 Evolutionary game theory

Contradicting the self-interest model, there is strong empirical evidence that humans tend to be cooperative even in controlled one-shot scenarios where being selfish would not damage their reputation (Cooper et al., 1996). Evolutionary game theory provides one explanation for why cooperation might be natural for people: Bravetti and Padilla (2018) argue that when $2R > T + S$, the population's average fitness grows in the event that both players cooperate compared to when one party exploits the other one. In competition for average fitness among populations, the ones with individuals favoring mutual cooperation outperform populations that on average incline to defection. (Bravetti and Padilla, 2018)

In a significant study on the evolution of cooperation, Axelrod and Hamilton (1981) conducted a competition in which famous game theorists were asked to submit strategies to a tournament of prisoner's dilemma. Every strategy would then play 200 rounds of prisoner's dilemma sequentially against each other and the one that got the highest cumulative score from all games would be the winner. As an input, each strategy had the history of what the opponent had played in the previous rounds of their match. Some of the strategies included cooperating until the opponent defects and after that only defecting (Friedman, table 6 label FR), defecting or cooperating randomly with 50% chance each (Random, table 6 label RAN) or cooperating on the first 11 rounds and then randomly cooperate with a probability of 10% less than how much the opponent cooperated on the first 10 rounds (Tullock, table 6 label TU).

The winner of the tournament was a strategy called tit-for-tat, which always cooperated on the first round and then simply copied what the opponent had done on the previous round. As table 6 shows, interestingly tit-for-tat never got more points than their opponent in a particular match. In spite of this, it yielded the highest cumulative points across all matches by cooperating with the strategies that relied on cooperation but also not being exploited by the strategies that relied on defection.

Prog.	Mean	Rank Point	No. of Wins	Rank Wins
TFT	504	1	0	15
T&C	500	2	11	2
NY	486	3	1	13.5
GR	482	4	4	6
SH	481	5	3	11.5
S&R	478	6	10	3.5
FR	473	7	6	8
DA	472	8	4	9.5
GR	401	9	5	9.5
DO	391	10	6	6
FE	328	11	12	3.5
JO	304	12	10	1
TU	301	13	6	6
NA	282	14	2	11.5
RAN	276	15	1	13.5

Table 6: Results from the tournament (Axelrod and Hamilton, 1981)

Recall that for maximizing payoffs in one-shot prisoner's dilemma, the game-theoretic optimal strategy is to always defect. The findings suggest that since the evolutionary environment is better represented by an iterated prisoner's dilemma between competing strategies, the advantage of cooperative or "nice" strategies has prevailed in the ecosystem.

3.3 Social preference models

The findings in evolutionary game theory (Axelrod and Hamilton, 1981; Bravetti and Padilla, 2018) suggest that cooperative equilibrium strategies emerge from repeated interactions between individuals in favorable conditions over time. The theory claims that this has resulted in genetic mutations making people intrinsically more cooperative. This intrinsic trait to cooperate helps in explaining the empirical evidence (Cooper et al., 1996) of people's tendency to cooperate even in one-shot interactions where cooperation is not rational.

The essence of social preferences is that they model intrinsic motivations, i. e. explain the human decision-making process and give insight into deeper underlying motivations behind the actions observed. For example, the inequality aversion model (Fehr and Schmidt, 1999) portrays people as having an innate dislike of inequally

distributed resources within their social groups. This can help in explaining societal phenomena such as how people tend to live in neighborhoods where the average income level is closer to their own. The following sections present some relevant social preference models and illustrate their mathematical formulation.

3.3.1 Inequality aversion

The inequality aversion model (Fehr and Schmidt, 1999) expects that people have an innate dislike of possessing or gaining a different amount of resources compared to the average in their reference group. For a group of size $n > 1$, the model can be represented mathematically as follows:

$$U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{i \neq j} max(x_j - x_i, 0) - \beta_i \frac{1}{n-1} \sum_{i \neq j} max(x_i - x_j, 0)$$

Where $\beta_i \leq \alpha_i$ and $\beta_i \in [0, 1[$

- $U_i(x)$ represents the utility function for player i.

- x_i represents the payoff for player i.

- x_j represents the payoff for the other player.

- α_i represents the aversion parameter for player i, indicating their sensitivity to inequality when the other player receives a higher payoff.

- β_i represents the aversion parameter for player i, indicating their sensitivity to inequality when they receive a higher payoff than the other player.

The essence of inequality aversion model is that there is aversion to both being at disadvantage and being at advantage in social exchanges. The base value of the utility function is still an individual's personal gain x_i and thus the model does not necessarily predict the highest utility to be the one without any inequality between group members but merely offsets some of the material gain with psychological aversion factors. The condition $\beta_i \leq \alpha_i$ states that in the model an individual is predicted to always be at least as much averse towards being at disadvantage as being at advantage and $0 \leq \beta_i < 1$ assumes that a person's aversion towards being

at advantage cannot be negative. If the parameter β was negative, that would mean that an individual would prefer being at advantage over equally distributed resources. Fehr and Schmidt (1999) acknowledge that there might be people with such aversion parameters but the condition is assumed for practical reasons.

3.3.2 Other preference models

In addition to inequality aversion, the field of behavioral economics has yielded several other social preference models that have shown reliable results in predicting outcomes of empiric economic games.

Homo moralis (Alger and Weibull, 2013) describes an individual that has a following utility function:

$$u_i(x, y) = (1 - \kappa_i) \cdot \pi(x, y) + \kappa_i \cdot \pi(x, x)$$

Where $\kappa_i \in [0, 1]$. The term "moralis" is based on the varying degrees of utility balancing between the individual's personal gain and the decision's benefit to the population as a whole. In the model $\pi(x, y)$ is the payoff that an individual gains by performing action x in the scenario. The altruistic factor $\pi(x, x)$ is the average payoff in the population if each agent in the population made the same choice as player i. According to the model, the individual makes a weighted choice based on their degree of κ that takes into consideration their balance between selfish and altruistic tendencies.

Other aspects of interaction may also be considered. Charness and Rabin (2002) presented mathematical formulations for concerns about social welfare and reciprocity. The social welfare model depicts individuals as caring about the welfare of others in their reference group and thus having inequality averse preferences. Reciprocity model assumes an individual to favor equal outcomes unless the agent they are interacting with has "misbehaved". This results in reciprocal behavior where the individual is not as concerned with the degree of equality towards the agent. Comparison shows that each of these models practically always outperform the self-interest model (Miettinen et al., 2020).

4 Human-AI strategic interactions

The purpose of this section is to review data from human-AI interaction studies that are based on the economic games introduced in section 3. Based on the literature and data, a framework providing an overview into the aspects of decision-making in strategic interactions is presented.

4.1 Cooperation with conflicting interests

The generalisation of AI systems has developed new challenges for how humans interact with machines. Previously, plenty of resources have been put into developing machine learning agents in order to cooperate with humans on mutual goals (Nikolaidis et al., 2015) or to outperform humans in zero-sum games such as chess or go (Zhang and Yu, 2020). In real life, most of cooperation does not occur in scenarios where the cooperation conditions would be this binary of completely aligned goals or winning an opponent in a competition. Many everyday interactions have actions taking place somewhere between zero-sum competition and full collaboration. For instance, business negotiations or taking care of one's family are positive-sum games: with productive cooperation each party in the interaction can gain positive value even though they still might have incentives to be selfish. Therefore, it is essential to study conditions where people interact in AI-augmented scenarios also in settings of conflicted interests. Social dilemmas model abstractly the scenarios of aligned and conflicting interests and are therefore an efficient platform to study human-AI interactions.

The factors shaping the cooperative decisions with AI agents are still unclear. Chu et al. (2023) found that by teaching human fairness models to AI agents, people who had economic interactions with such agents found the fair types more warm, intelligent, likable and safe compared to AI that was not trained to model human behavior. However, experiments in economic games have shown that even when anticipating cooperation from AI agents, humans tend to not reciprocate the benevolence of non-human agents (Karpus et al., 2021). An important question becomes how AI agents can be trained and altered to be cooperative while also avoiding the agents getting exploited by humans.

4.2 Social preferences in AI-augmented scenarios

Social preferences have been studied extensively for human-human interactions (Fehr and Schmidt, 1999; Alger and Weibull, 2013; Miettinen et al., 2020). However, it is unsettled in scientific literature how social preferences apply when humans interact in AI-augmented scenarios. AI could be implemented to strategic interactions as autonomous agents or humans utilizing AI tools. Therefore, humans interacting with varying forms of AIs might induce social preferences differently. The results in social dilemma games from interactions between humans and autonomous AI agents in online settings seem to indicate lowered social preferences towards AI (Karpus et al., 2021).

	Prisoner's dilemma	Chicken game	Stag Hunt
Cooperation rate vs. Human	0.49*	0.69*	0.89
Cooperation rate vs. AI	0.36*	0.56*	0.80
Expected cooperation vs. human	0.59	0.67	0.78
Expected cooperation vs. AI	0.52	0.70	0.79
Cooperation rate vs. human when expected cooperation	0.71*	0.73*	0.98*
Cooperation rate vs. AI when expected cooperation	0.54*	0.57*	0.91*

Table 7: Data for human players from (Karpus et al., 2021). * denotes statistical significant difference within comparison where $p < 0.05$.

The data from table 7 provides different explanations for human cooperation with AI agents. Because the action of defection might be motivated by exploitation when expecting cooperation and the fear of defection from the opposing player, cooperation rate alone is not sufficient to study social preferences arising in this scenario. Due to the payoff structure of prisoner's dilemma and chicken game, cooperation rate in those games reveal self-interested tendencies when the participant expects the opposing player to cooperate. Conversely, defecting in response to expected defection would indicate aversion towards being at disadvantage.

For stag hunt, defection is not in the interests of selfish players who believe the other participant to cooperate as the action would yield lower payoff than mutual

cooperation. On the other hand, stag hunt reveals possible distrust between players as a player expecting defection would prefer to defect to get a guaranteed medium payoff. Another possible explanation for defection in stag hunt could be willingness to outperform one's opponent when expecting cooperation. However according to Karpus et al. (2021), in stag hunt the significant difference in behavior when expected cooperation is explainable by lower confidence levels in expected cooperation against AI agents.

The data from prisoner's dilemma and chicken game show that although participants expected cooperation on closely equal rate from humans and AI agents, the cooperation rate is significantly lower on both games against AI agents. This suggests that social preferences towards equal outcomes are strictly lower for humans who play against AI agents since the expected benevolence of AI agents is not reciprocated as often.

	Trust game 1st player	Trust game 2nd player
Cooperation rate vs. Human	0.74	0.75*
Cooperation rate vs. AI	0.78	0.34*
Expected cooperation vs. human	0.56	0.79
Expected cooperation vs. AI	0.55	0.83
Cooperation rate vs. human when expected cooperation	0.91	0.80*
Cooperation rate vs. AI when expected cooperation	0.86	0.35*

Table 8: Data for human players from (Karpus et al., 2021). * denotes statistical significant difference within comparison where $p < 0.05$.

Due to the sequential nature of trust games, the players receive asymmetric information about the other player's move with respect to their own decision-making. Player 2 acts in full knowledge of what the first player did and therefore defecting as the second player may only be seen as an act of exploitation towards the first player who appeared cooperative. As table 8 illustrates, player 1 participants show no significant difference between cooperating against AI agents and humans. The difference between expecting the second player to reciprocate the cooperative choice was not significant either. When expecting cooperation, most of player 1 participants

cooperated regardless of the non-human nature of the opposing player. Moreover, the difference between cooperation rates and expected cooperation from another player can again be explained by varying levels of confidence in the first player predictions (Karpus et al., 2021). Therefore a possible reason for the effect is that some of the players still cooperate despite expecting defection in hopes of being wrong with their prediction. For the second player participants, the data shows significant difference between cooperation rates against human and AI agents. This suggests much lower social preferences for reciprocity and equality towards AI compared to human opponents.

Conducting a simple reanalyzis for prisoner's dilemma and chicken game cooperation rates with the social preference model homo moralis (Alger and Weibull, 2013) yields the degree of morality an individual should have to cooperate.

$$u_i(x, y) = (1 - \kappa_i) \cdot \pi(x, y) + \kappa_i \cdot \pi(x, x)$$

Where $\kappa_i \in [0, 1]$, $\pi(x, y)$ is the personal payoff for player i for action x, $\pi(x, x)$ is the average payoff in the reference group if everyone performed action x in the scenario. When applying this model for 2x2 symmetric matrix games with payoffs T, R, P, S the decision to cooperate can be analyzed followingly. For simplicity, the calculations only consider situations where the opponent is expected to cooperate.

$U_i(C) > U_i(D) \rightarrow u_i(C, C) > u_i(D, C)$

$\rightarrow (1 - \kappa_i) \cdot R + \kappa_i \cdot R > (1 - \kappa_i) \cdot T + \kappa_i \cdot P$

Solving for κ_i, homo moralis (Alger and Weibull, 2013) predicts that in table 7 prisoner's dilemma with payoffs $T = 100$, $R = 70$,$P = 30$, $S = 0$ (Karpus et al., 2021), the participants who cooperated would have to possess a morality value of $\kappa > 0.429$. For chicken game with payoffs $T = 100$, $R = 70$,$P = 0$, $S = 20$ (Karpus et al., 2021), the morality value would have to be $\kappa > 0.3$. In table 7, the recorded cooperation rates are higher in the chicken game than in the prisoner's dilemma as indicated by the differences in κ threshold. As past research suggests (Miettinen et al., 2020), analyzing payoff structures with homo moralis seems to have credibility in predicting cooperation rates for strategic interactions between humans.

Recall that in table 7, cooperation rates when expected cooperation in prisoner's dilemma were significantly lower against AI agents than against human opponents.

However, homo moralis (Alger and Weibull, 2013) does not address the possibility of interacting with non-human entities in its mathematical formulation. Thus with equal payoffs, the predicted κ threshold should be the same regardless of the opponent's identity. The data still shows that participants cooperate less with AI agents. This suggests that social preference models are unable to predict the change in attitudes towards non-human agents in strategic interactions. Assuming that the preferences were identical against both humans and AI agents, the differences in cooperation rates would be insignificant.

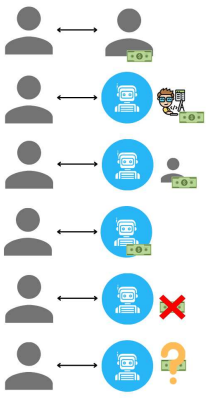
4.3 Human beneficiary behind the machine

In a meta-analysis of online anonymous human-machine economic interactions, von Schenk et al. (2023) found that approaches on how the machines are introduced to the participants vary, which might influence the results. Their experimental data suggests that people have different social preferences towards machines depending on where the payoffs gained by the machine are believed to end up.

Table 9: Overview of results from von Schenk et al. (2023).

Opponent's introduction	**Type 1** mildly altruistic/ selfish	**Type 2** spiteful	**Type 3** aheadness averse/ positively reciprocal
Control group: Fellow human	16.11%	10.74%	73.15%
Programmer behind Machine	27.33%	20.67%	52.00%
Token Player behind Machine	26.67%	15.33%	58.00%
Machine Earns	36.00%	41.33%	22.67%
Nobody Earns	31.33%	55.33%	13.33%
No Information	30.67%	43.33%	26.00%

Based on the recorded behavior on social preferences, participants were divided into three groups. The first group was "mildly altruistic/selfish" who showed little inequality aversion nor reciprocal preferences. The "spiteful" group had strong preference of being ahead in payoffs and disliked being at disadvantage. Their degree of reciprocity was not significant either. The third group was "aheadness-averse/positively reciprocal". This group disliked advantageous outcomes in their favor and reciprocated kind acts towards their co-player. The group could be described as "altruistic" or "pro-social".



The study (von Schenk et al., 2023) showed that when the participants believed no one would receive the money or had no information about the receiver, they were more likely choose an exploitative option and showed weak social preferences for equality. Conversely, when participants were told that the money received by the machine agent is transferred to the programmer or another random participant, the results showed heightened cooperative preferences. Furthermore, both positive and negative reciprocity were reduced towards machines operating without an implied human partner. In post-analysis questionnaire, a major part of the no-information group believed that no one received the money gained by the AI. However, these beliefs tended to form self-favorably to justify the participants' own exploitative actions as non-involved participants believed much less in no one receiving the payoffs (von Schenk et al., 2023).

The findings (von Schenk et al., 2023) demonstrate that in online settings, people frame their social preferences based on who they think is the actual receiver of the payoffs in the economic games. If the receiver of the payoffs is non-human, the cooperation rate is lower. This may also help to explain findings where the ambiguity of the agent's non-human identity increases cooperation rates (Ishowo-Oloko et al., 2019). However, even when participants knew that the payoffs earned by the machine were given to humans, the altruistic tendencies decreased. Still introducing human beneficiary behind the machine strengthens social preferences compared to no information or the non-human agent earning the payoffs. Thus, future research on how autonomous agents in real life are perceived in terms of possible human beneficiary behind the machine is necessary to determine social preferences more effectively.

4.4 Achieving human-level cooperation with AI agents

Combining expert algorithms and reinforcement learning, AI agents have shown to elicit cooperative behavior when interacting with humans online in sequentially played social dilemmas (Crandall et al., 2018). The S++ algorithm for instance relies on being cooperative and forgiving some instances of defection but still punishing prolonged defection. Moreover, introducing an element of non-binding communication resulted in elevated cooperation rates when the algorithm interacted with humans (Crandall et al., 2018). However, a problem with this approach is that interacting

online people cannot tell whether they are playing with an AI agent or actual human unless they are being provided the information explicitly. As illustrated in the data from table 9, this might form a challenge when translating the success of these cooperative algorithms to real life interactions where the non-human nature of the machine is unambiguous.

The fact that people cooperate less with explicit machine agents (von Schenk et al., 2023) creates a problem called transparency-efficiency tradeoff (Ishowo-Oloko et al., 2019) where informing people accurately about their partners non-human nature reduces the observed cooperation rates. Furthermore, brain regions associated with considering other people's intentions and thoughts show reduced activation in social dilemmas when humans knowingly are playing against non-human opponents (Rilling et al., 2002). This illustrates that the subjects tend not to perceive machines as sentient entities who possess intentions. This might also depend on the subject's cultural background as eastern cultures tend more often to attribute machines to have a spirit or a soul (De Freitas et al., 2023). In the western cultures the belief that machines are not sentinent or possess a soul is still prevalent. Therefore an important question becomes how to avoid the decrease of social preferences towards confirmed non-human agents.

A major factor influencing human attitudes towards AI agents is how emotionless the machines are perceived to be (De Freitas et al., 2023). Moreover, humans tend to consider AI agents socially as out-group members (De Freitas et al., 2023). Thus, research indicates that introducing emotional expressions and cultural characteristics to machine agents elicits cooperation (de Melo and Terada, 2019). Contrary to playing against machine agents without implemented human-like features nor human beneficiaries (Karpus et al., 2021; von Schenk et al., 2023), anthrpomorphizing machine interactions with avatars that expressed joy for mutual cooperation and sadness for exploitation resulted in almost equal cooperation rate regardless whether the participant was told that the opponent was human or machine agent (de Melo and Terada, 2019).

4.5 Formation of the decision to cooperate

The social preference models (Fehr and Schmidt, 1999; Alger and Weibull, 2013) frame the human decision of cooperating as a balance between self-interest and

individual altruistic tendencies. They are thus only able to grasp a marginal aspect of the process on how the choice of cooperation is formed. For instance, homo moralis (Alger and Weibull, 2013) describes altruistic factor as average payoff for the individual's social group expecting that everyone in the group behaved identically to them. Decision is then made based on the degree of morality with comparing utilities according to an individual's value of κ balancing self-interest and altruism.

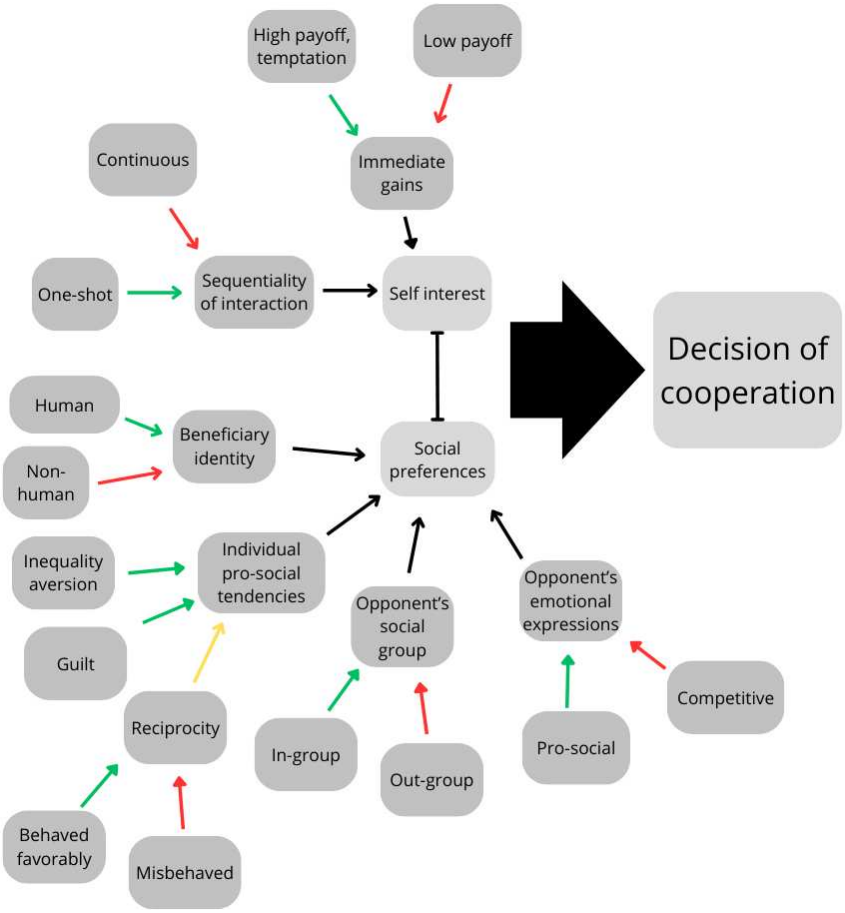


Figure 2: A framework of human decision-making in strategic interactions. Green and red arrows correspond to positive and negative reinforcement towards self-interest or social preferences

As evident by results from interacting with machine agents, the cooperative decisions in economic games tend not only to depend on the degree of individual's pro-social tendencies but also on context-dependent aspects such as what is the identity of the opponent (Ishowo-Oloko et al., 2019), who is the actual beneficiary of the interaction (von Schenk et al., 2023) and how relatable the opponent is perceived to be (de Melo and Terada, 2019).

Based on the literature of economic games presented in this thesis, figure 2 introduces a framework for human decision-making. The structure of the framework originates from behavioral economic mathematical models (Alger and Weibull, 2013; Fehr and Schmidt, 1999; Charness and Rabin, 2002) and is elaborated with findings of evolutionary game theory (Axelrod and Hamilton, 1981), empirical studies on decision-making in interactions with machine agents (Karpus et al., 2021; Ishowo-Oloko et al., 2019; de Melo and Terada, 2019; von Schenk et al., 2023) and psychological literature (De Freitas et al., 2023). The framework combines elements from mathematical models and addresses that contrary to social preference models, there are several psychological factors in the interaction that correlate to how likely participants are to cooperate. As shown by findings from evolutionary game theory (Axelrod and Hamilton, 1981), self interest factor is not only dependent on the immediate payoffs in the scenario but also on the sequentiality of the interaction. The decision of cooperation in this model can be thought of as a comparison between self-interest and social preference factors. The framework predicts that if the degree of social preferences proportionally outweights self-interested component, the decision will be cooperative.

A limitation to this framework's predictive capability is that the human decision to cooperate in real life cannot be represented by a calculation where each factor would be considered simultaneously with equal weight. For instance, even if the opponent had shown pro-social emotional expressions (de Melo and Terada, 2019) and behaved favorably (Charness and Rabin, 2002), introducing novel information that no one would receive the payoffs would still likely lower the cooperation rates (von Schenk et al., 2023). Therefore one limitation for the framework is that it does not consider the possible recency effects where new updated information would have a larger impact in the decision-making. Moreover, another limitation is that the framework only considers context-dependent components of the interaction. On individual level there are also other factors such as the mental state that might affect the decisions. Still, an important advantage is that the framework considers interactions with non-human entities, which has previously been ignored in models of human decision-making. The framework could be used to ideate mathematical formulations to develop more accurate behavioral game theoretic tools.

5 Conclusion

This thesis aimed to present the current research on human cooperation and human-AI strategic interactions. With means of behavioral economics and social preference models, literature on economic games and psychological factors was reviewed to analyze what makes people cooperate when they are interacting with human and non-human entities. In addition to reviewing literature, a framework to evaluate context-dependent aspects of cooperation in human and human-AI interactions was presented. The decision of cooperation was essentially identified to consist of a balance between self-interest and social preferences. Additionally to theories traditionally associated with affecting economic decision-making between humans, concepts such as beneficiary identity or opponent's perceived relatability were recognized to affect the cooperation in AI-augmented scenarios.

From a societal point of view, it is important to notice that manipulating human cooperation in strategic interactions may be motivated by economic or political benefit for malevolent actors. For instance, as intelligent AI agents start navigating online platforms, their enhanced capability to elicit cooperation could be abused for malign purposes. Therefore, it is important to be prepared to tackle this issue with means of legislation and educating the public on AI agents' possible manipulative capabilities. Nevertheless, creating and maintaining cooperative relationships between benevolent actors is crucial for a functional society in order to guarantee a safe and productive environment for its inhabitants. This cooperation has to be sustained even when the interactions involve AI agents or AI-augmented people.

In addition to the limitations of the framework introduced in section 4, there are some general limitations for the thesis. Firstly, the studies covered that assess human cooperation through economic games are conducted in controlled conditions. Thus, they might not fully capture the complex nature of real-life strategic interactions. Secondly, the element of novelty that is often present when interacting with AI systems in study settings might influence the results. As the AI development progresses further and humans learn more about AI systems, the human behavior and cooperative tendencies may change differently in strategic interactions. Lastly, new papers are published continuously and some relevant studies may not be covered.

Based on the findings in the thesis, many directions for future research exist. As the beneficiary identity was identified to have an important effect on social preferences,

assessing how different AI agents such as self-driving cars are perceived in real-life interactions in terms of their human beneficiaries is important. Moreover, the current social preference models were shown to be unsuitable in predicting cooperation with non-human agents. Therefore, the field of behavioral game theory would make use of more accurate mathematical models that also consider strategic interactions in AI-augmented scenarios.

References

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302. https://doi.org/10.3982/ECTA10637.

Alkatheiri, M. S. (2022). Artificial intelligence assisted improved human-computer interactions for computer systems. *Computers and Electrical Engineering*, 101:107950. https://doi.org/10.1016/j.compeleceng.2022.107950.

Aschenbrenner, L. (2024). Introduction - situational awareness: The decade ahead. Accessed: 01.08.2024. Available: https://situational-awareness.ai/.

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396. https://doi.org/10.1126/science.7466396.

Boksberger, P. E. and Melsen, L. (2011). Perceived value: a critical examination of definitions, concepts and measures for the service industry. *Journal of services marketing.*, 25(3). https://doi.org/10.1108/08876041111129209.

Bravetti, A. and Padilla, P. (2018). An optimal strategy to solve the prisoner's dilemma. *Scientific reports*, 8(1):1948. https://doi.org/10.1038/s41598-018-20426-w.

Camerer, C. F. (1997). Progress in behavioral game theory. *Journal of Economic Perspectives*, 11(4):167–188. https://doi.org/10.1257/jep.11.4.167.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The quarterly journal of economics*, 117(3):817–869. https://doi.org/10.1162/003355302760193904.

Chu, Y., Li, J., and Xu, J. (2023). How is the ai perceived when it behaves (un)fairly? In Degen, H. and Ntoa, S., editors, *Artificial Intelligence in HCI*, pages 421–430, Cham. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_25.

Cooper, R., DeJong, D. V., Forsythe, R., and Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, 12(2):187–218. https://doi.org/10.1006/game.1996.0013.

Crandall, J. W., Oudah, M., Chenlinangjia, T., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J., Cebrian, M., Shariff, A. F., Goodrich, M. A., and Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, 9. https://doi.org/10.1038/s41467-017-02597-8.

De Freitas, J., Agarwal, S., Schmitt, B., and Haslam, N. (2023). Psychological factors underlying attitudes toward ai tools. *Nature Human Behaviour*, 7(11):1845–1854. https://doi.org/10.1038/s41562-023-01734-2.

de Melo, C. M. and Terada, K. (2019). Cooperation with autonomous machines through culture and emotion. *PLoS ONE*, 14. https://doi.org/10.1371/journal.pone.0224758.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868. https://doi.org/10.1162/003355399556151.

Hacker, P. (2018). Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under eu law. *Common market law review*, 55(4). https://doi.org/10.54648/cola2018095.

Hu, Y., Zhang, S., and Dang, T. (2024). Exploring large-scale language models to evaluate eeg-based multimodal data for mental health. https://doi.org/10.48550/arXiv.2408.07313.

Ishowo-Oloko, F., Bonnefon, J., Soroye, Z., Crandall, J. W., Rahwan, I., and Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1:517 – 521. https://doi.org/10.1038/s42256-019-0113-5.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of economic psychology*, 32(5):865–889. https://doi.org/10.1016/j.joep.2011.05.007.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292. https://doi.org/10.1142/9789814417358_0006.

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., and Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent ai. *Iscience*, 24(6). https://doi.org/10.1016/j.isci.2021.102679.

Kurvers, R. H., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., and Bahrami, B. (2021). Strategic disinformation outperforms honesty in competition for social influence. *Iscience*, 24(12). https://doi.org/10.1016/j.isci.2021.103505.

Kusner, M. J. and Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, 578(7793):34–36. https://doi.org/10.1038/d41586-020-00274-3.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133. https://doi.org/10.1007/BF02478259.

Miettinen, T., Kosfeld, M., Fehr, E., and Weibull, J. (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *Journal of Economic Behavior Organization*, 173:1–25. https://doi.org/10.1016/j.jebo.2020.02.018.

Mullainathan, S. and Thaler, R. H. (2000). Behavioral economics. https://doi.org/10.3386/w7948.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., and Forghani, R. (2020). Brief history of artificial intelligence. *Neuroimaging Clinics of North America*, 30(4):393–399. https://doi.org/10.1016/j.nic.2020.07.004.

Nikolaidis, S., Ramakrishnan, R., Gu, K., and Shah, J. (2015). Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 189–196, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2696454.2696455.

Nowak, M. A. et al. (2012). Evolving cooperation. *Journal of theoretical biology*, 299(0):1–8. https://doi.org/10.1016/j.jtbi.2012.01.014.

Nowak, M. A., Page, K. M., and Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, 289(5485):1773–1775. https://doi.org/10.1126/science.289.5485.1773.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192. https://doi.org/10.1126/science.adh2586.

Osborne, M. J. (1994). *A course in game theory*. MIT Press. https://sites.math.rutgers.edu/~zeilberg/EM20/OsborneRubinsteinMasterpiece.pdf.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., and Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, 35(2):395–405. https://doi.org/10.1016/S0896-6273(02)00755-9.

Sadiku, M. N. O. and Musa, S. M. (2021). *Augmented Intelligence*, pages 191–199. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-77584-1_15.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., and Teller, A. (2016). Artificial intelligence and life in 2030. Stanford University. Available: http://ai100.stanford.edu/2016-report.

von Schenk, A., Klockmann, V., and Köbis, N. (2023). Social preferences toward humans and machines: A systematic experiment on the role of machine payoffs. *Perspectives on Psychological Science*, 0(0). https://doi.org/10.1177/17456916231194949.

Zhang, H. and Yu, T. (2020). *AlphaZero. In: Deep Reinforcement Learning.*, pages 391–415. Springer, Singapore. https://doi.org/10.1007/978-981-15-4095-0_15.