University of Zurich[UZH]

Department of Sociology

Master Thesis

Evading the Algorithm: The Increased Propensity for Tax Evasion in Human-Computer Interactions

An Empirical Analysis of Human Strategic Decision-Making Processes in a Tax Evasion Inspection Game

Author:	**Nico Mutzner**
Student ID #.:	21-709-811
Email:	nico.mutzner@bluewin.ch

Module Name:	Master Thesis
Module ID:	06SM240-MA
Semester/Year:	FS23
Supervisor/Lecturer:	Dr. Prof. Heiko Rauhut
Submission Date:	15.05.2023

Abstract

Today's modern world is characterized by an increasing shift from human-to-human interaction towards human-computer-interaction (HCI). With the implementation of artificial agents as inspectors, as can be seen in today’s airports, supermarkets or most recently within the context of the COVID-19 pandemic, our everyday life is progressively shaped around interacting with automated agents. While various studies have looked at cooperative strategic interaction between humans, little is known about how HCI affects humans and their non-cooperative decision-making. Therefore, a deeper understanding of the factors influencing strategic decision-making processes within HCI situations, and how perceptions of automated agents’ capabilities might influence these decisions is required. This gap is addressed by extending a non-cooperative inspection game experiment with a tax evasion frame, implementing automated agents as inspectors. Hereby a within-subject design is used to investigate (1) how HCI differs from human-to-human interactions in this context and (2) how the complexity and perceived capabilities of automated agents affect human decision-making. The results indicate significant differences in decisions to evade taxes, with participants more likely to evade taxes when being inspected by automated agents compared to humans. Participants were also less likely to evade taxes when playing against an automated agent described to be a complex AI compared to an automated agent described to be a simple algorithm once they had experienced different agents.

1 Introduction

We see an ever-increasing amount of technology entering our everyday lives, with technological implementations finding their way into almost all aspects of our social realities. This pervasiveness of technology necessarily comes with an increase in exposure to technology, which consequently leads to more frequent interaction patterns between humans and automated agents (AAs). AAs are a physical technology, often mechanized or computerized, designed to minimize the need for human intervention in a defined environment (Kaber, 2018). Within the scope of this study, AAs serve as substitute agents, supplanting human agents in particular functions and altering the dynamics of the interaction paradigm. While interactions between humans and AAs might be deemed as simple or straightforward at first glance, the public as well as researchers have found such interactions to be much more complex. For the public, the attention is often focused on the impact of the digital life that we live today, characterized by our shift towards a technologically enabled online life, the prevalent use of social media and an overall reliance on technology for many everyday tasks. Further, implementations of artificial intelligence (AI) have produced a vivid image of technological advance, one that comes with great promises and equally great pitfalls. We have seen such duality in the great promises that arose with AI implementations in fields such as medicine (He et al., 2019) or self-driving cars (Sestino et al., 2022), but also in the rise of critical discussions about AI's shortcomings, such as facial recognition biases (Leslie, 2020; Lohr, 2022). Very recent discussions about the new natural language model GPT-3 & GPT-4, and it's application in ChatGPT (OpenAI, 2023, 2022), have raised questions about how we interact with machines and technology (Roose, 2023; Stokel-Walker, 2023). Our utilization of technology not only shapes our decision-making processes, but crucially, our perceptions of technology can significantly impact our strategic engagements with these systems. Researchers have recognized the importance of analyzing this interaction with machines and identified the need to find appropriate theories and experiment which can explain the differences between human-to-human interaction and human-computer interaction (De Melo et al., 2016).

This paper aims to elucidate these differences between human-to-human and human-computer interaction in the context of strategic decision-making. While there has been increasing interest in studying these relations in experiments such as the prisoners' dilemma or economic games like dictator games, ultimatum games, negotiation games, and public goods games (De Melo et al., 2016; Kiesler et al., 1996; Lee et al., 2021; Nielsen et al., 2022; Weiss et al., 2020), the incorporation of AAs in the inspection game remains less explored. The inspection game is a non-cooperative economic game with a mixed-strategy equilibrium, meaning that there is no pure strategy to follow, and players have to rely on strategic decision-making. Such crime detection games have been argued to well represent social interaction effects (Falk and Fischbacher 2002) and lend themselves well to be played in different frameworks, such as the tax evasion framework chosen for this experiment. Traditionally, the inspector within this economic game has been another human player. However, in this study, we implement an AA as the inspector, manipulating who the participants believe they are playing against. This leads to the central research question of this paper: How does strategic decision-making differ between human-to-human and human-computer interaction when placed in a non-cooperative strategic setting, and does the complexity of the computer affect potential differences? By having participants think they are playing against different agents - a human, a simple AA, or a complex AA - we can identify how the deployment of different agents impacts participants' strategic decision-making. Therefore, this experiment reflects the increasing use of computer systems to automate previously human-controlled functions, specifically in controlling deviant behavior. By identifying differences in decision-making, we can better understand how these changes affect strategic decision-making processes and their consequences on norm-deviating behavior.

Results from 300 Participants in an online experiment reveal distinct variations in tax evasion behavior when participants are put against perceived human players and AAs. Both linear and mixed effects logistic regression results indicate significant differences in interactions with human agents as opposed to automated ones, as well as between perceived simple and complex AAs in later rounds. We further find significant round effects, where participants evasion probabilities would either reduce or increase over the 15 rounds played dependent on agent type. Getting caught also affected participants' decisions, greatly improving the probability of trying to evade taxes in the next subsequent round. However, the effects of the perceived agent type remain significant, even when considering these and other confounding variables. The results indicate clear differences in strategic behavior that is dependent on who people think they are playing against, with human opponents eliciting lower norm-deviating behavior in the form of evading taxes. Contrary to findings in previous studies, this effect does not seem to be mediated by either technical affinity or tax attitudes. Higher evasion probabilities were also found to be affected by attitudes towards the wider implementation of AAs in people's lives, with people disagreeing with such a wider implementation showing higher tax evasion rates in the human treatment compared to the complex AA treatment. These findings contribute to our understanding of the implications of AA implementations in control and inspection functions. Overall, the results can shed light on the complexities of strategic human-computer interaction and inform more effective strategies for deploying automated systems in roles traditionally performed by humans.

2 Literature Review

2.1 Economic Games

Becker (1968) first introduced an economic approach towards deviance and crime in an attempt to develop "optimal public and private policies to combat illegal behavior" (p. 207). He employs variables for diverse expenditures, losses, and costs with which to analyze and calculate the efficiency of measures to combat illegal behavior and reduce social loss. This approach provides an insightful look at how crime can be quantified and tied to the resources used to combat crime. The original version of the economic inspection game can be found in Dresher's (1962) work which focused on the strategic settings of a smuggler and an inspector. In a similar fashion, Maschler (1966) used the inspection game to formulate a non-constant sum game in which an inspector and a violator enter an agreement in which the inspector is allowed to inspect a fixed number of times while the violator can choose to violate one time throughout *n* rounds. Both these early versions of the inspection game employ a limited number of violations and inspections which can be useful when considering certain real-life occurrences with limited inspections, for example in the case of arms control agreements. Yet, for situations where criminal violations and inspections are only limited by costs and risk factors, it makes sense to place no such constraints. Consequently, removing such constraints also shifts the focus away from the previous approaches, which rely heavily on a more economic model of the inspection game concept, and instead moves it more towards a sociological and criminological understanding of criminal behavior. In this way, this study positions itself along the research of Tsebelis (1989, 1990), who has introduced the necessity of looking at crime from a game theoretic perspective, which reflects the mixed strategy equilibrium employed by rational opponents compared to probabilistic measures employed within decision theory. Yet, as Bianco, Ordeshook, and Tsebelis (1990) have pointed out, the one-shot nature of the experiment employed by Tsebelis (1989, 1990) led to wrong representations of the actual phenomenon of crime, where decisions by citizens and police officers are made continuously over time. To this end, authors such as Andreozzi (2004) have employed a sequential simultaneous version of the inspection game, where decisions are made over several rounds, and decisions are made by both players at the same time. This was further extended by Rauhut and Jud (2014), who focus on a social norms approach towards detection and punishment, where inspectee's are labeled as unknown norm violators. In contrast to previous iterations of the inspection game, in this version the action to inspect or control is associated with a cost - but can also generate a reward upon successful detection of a crime. By employing these additional factors, they produce a model in which there is no equilibrium in pure strategies, and participants are forced to strategize to reach a decision within each round. This discoordination situation is critical within this proposed study, as it is reliant on the participants having to strategize and not choose a predefined optimal strategy, which in turn supports the focus on differing strategies against different agents.

Building upon this mixed strategy foundation within the inspection game, we can then place it within the frame of tax evasion. Tax evasion experiments have been employed for some time but have recently seen increasing attention within academic literature. While they are often used to address issues in tax administration and compliance, they are fundamentally based upon the economics of crime framework of Becker (1968), and therefore share the strategic economic decision foundation found in inspection games (Mascagni, 2018). Determining factors of tax compliance within these experiments are both based on economic models (Beck et al., 1991) as well as social determinants such as norms and ethics (Blumenthal et al., 2001; Torgler, 2002, 2007). Importantly, the tax evasion framework allows the identification of causal relationships with the introduction of independent variables (Spicer & Thomas, 1982). By keeping the other independent variables constant, one can introduce an independent variable of interest to evaluate changes within the tax evasion behavior of participants. In line with deviant behavior in the wider application of inspection games, tax evasion decisions in experiments come with a moral and emotional drawback of cheating behavior, which can be placed in a social paradigm, in contrast to purely economically focused activities such as gambling (Baldry, 1986; Coricelli et al., 2010). It is possible to influence this morality aspect as well as the wider psychological aspects underlying the decision to evade taxes by manipulating external factors within the experimental setting (Webley & Halstead, 1986). In essence, experimental literature has, much like the inspection game literature, recognized the fact that purely economic models of utility are not enough to explain human decision-making behavior within these situations, and a myriad of social factors have to be considered to explain the phenomenon (Alm, 2012; Lefebvre et al., 2015; Mascagni, 2018). Even with the inclusion of the wider social factors, most studies both in the tax literature as well as the inspection literature have focused mostly on the taxpayer or inspectee’s themselves, framed within the economic constraints, and have not expanded their considerations to the agents doing the inspection. This study addresses this gap by employing different agents as inspectors, more specifically by including AAs. This helps to demonstrate the impact strategic interaction agent constellations can have on human decision-making, informing our fundamental understanding of strategic decision-making when interacting with different agents.

2.2 Human-Computer Interaction

Much of the study of social decision-making and decision-making behavior is based upon the notion of human agents being placed in specific interaction settings, as exemplified in the inspection game and tax evasion literature. However, what if the agents are not humans, but instead machines? Nass, Steuer, and Tauber (1994) have addressed this idea in their paper “Computers as social actors”, where they attempted to prove that human computer interaction is based on social foundations, and experiments could therefore elicit social behaviors from participants when they are paired with AAs. They tested this in an experiment with a student population that participated in a computer tutoring session and found that participants apply social characteristics to the computers, including social norms, notions of self and others, gender, and social response. Some of these results were replicated in later studies, such as applying gender norms onto computers due to a gendered voice output (Nass et al., 1997), reacting to emotional displays of virtual agents (de Melo et al., 2014), and categorizing computers as in-group or out-group (Eyssel & Kuchenbrandt, 2012). Computers can also be seen as teammates, where humans that are teamed up with computers will behave in a similar fashion than when interacting with a human, even showing higher conformity and trust with computers (Nass et al., 1996; Robinette et al., 2016; Salem et al., 2015). Nass and Moon (2000) explain the existence of these social attributions onto computers on the basis of Langers (1992) concept of mindlessness. Mindlessness can be described as a state in which a person relies heavily on categories and distinctions formed in the past, which can override current aspects of a situation. Nass and Moon (2000) argue that such a process also takes place when humans interact with a computer, where social scripts are activated which in turn lead to the social nature of the interaction. Further studies confirmed that when humans are placed in an experimental game with computers, they attribute intentionality, desire as well as mental states to computers (Gallagher et al., 2002; Krach et al., 2008). This breadth of studies exemplifies that there is an inherent and active social nature with which we interact with computers, even though we may not be fully aware of it. Therefore, research into these areas of interaction is necessary to see how such notions can affect the decisions humans make when placed with or against a computer.

Historically, the inclusion of AAs to test such considerations has been sparce. For the inspection game, AAs have mostly been used as automated tools to simulate decisions of rational learning models (Rauhut, 2015), or multi-agent systems used for automated negotiation (Radu, 2015). Yet, as human-machine interaction becomes more prevalent and relevant, it becomes necessary to include AAs not only as a simulation tool, but also have them included as an active player in the interaction scenarios. One of the first examples of including computers in an experiment was Kiesler, Sproull, and Waters (1996), who proved results by Nass et al. (1994) by showing that humans show characteristics of social interaction when interacting & cooperating with technology, and follow social rules when placed in a prisoner’s dilemma with computers. Participants proposed cooperation with computers similarly to human counterparts but would do less so if the computer was more human-like. One of the pioneering studies that specifically investigated the differences in strategic interactions between humans and computers within a strategic setting was conducted by (De Melo et al., 2016). Participants played a public goods game, a dictator game as well as an ultimatum game, with both human and computer treatments. Firstly, they concluded that participants showed social considerations of their computer counterpart by allocating money into the shared pool in public goods games, as well as extending non-zero offers in ultimatum and dictator games. These decisions of contributing to computers or trusting computers, even though it goes against the rational strategy, were later further replicated in three different studies (Nielsen et al., 2022; Schniter et al., 2020; Weiss et al., 2020). Critically, De Melo et al. (2016) also found that participants were more likely to cheat and exploit computers and AI compared to humans. This tendency was reproduced in other studies, where participants did show trust in AAs but were more likely to exploit them (Karpus et al., 2021) as well as people reporting more dishonestly towards AA’s compared to humans in a coin toss task (Maréchal et al., 2020). Therefore, while people treat AA’s socially, they still tend to be dishonest and exploit them more than humans. This leads to the first hypothesis:

> *H1: Participants are more likely to evade taxes if they perceive the inspector to be an automated agent compared to a human inspector*

Yet, how much this exploitation and dishonesty takes place can depend on the characteristics of the AA. Focusing on how the mind of an AA is perceived, Lee, Lucas, and Gratch (2021) looked at how the modelling of an agent along agency and patiency parameters can influence human decision-making. They had participants play a dictator game, an ultimatum game as well as a negotiation game with manipulated perceptions of artificial agents. They found that altering agency and patiency does induce changes within the outcomes of the game, suggesting people perceive such attributions and change their strategy accordingly. Further, higher complexity of the algorithm can elicit higher cooperation (Crandall et al., 2018). However, the perceived complexity of an AA does not necessarily have to correspond with its actual complexity, but can be based solely on the agent's perceived characteristics. For example, an agent which is believed to be more altruistic/selfish will elicit different strategic decisions from participants in economic games (Daylamani-Zad & Angelides, 2021). This perceived complexity of an agent can be induced through a description of the agent. Langer et al. (2022) have shown that terminology with which an AA is described, including terminology such as "Artificial Intelligence" and "Algorithm", produce differences in participants perceptions of fairness, trust and justice. Considering the mindlessness concept by Langer (1992), participants can also be more likely to fall back on established social scripts and risk estimations when the opposing agent is perceived to be more closely aligned with a human agent. In the setting of a non-cooperative decision-making game, variances in strategic choices can therefore be observed when participants interact with different types of Autonomous Agents (AAs). Notably, participants often attribute enhanced capabilities to what they perceive as more complex AAs, and estimate a higher risk of detection, therefore reducing their evasion behavior in such situations. This brings us to our second hypothesis:

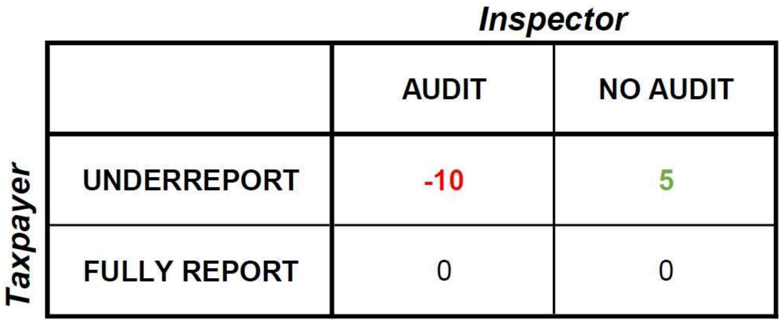
> *H2: Participants are more likely to evade taxes when they perceive the automated agent to be a simple algorithm compared to an automated agent described as a complex artificial Intelligence*

3 Methods

3.1 The Inspection Game

Participants engaged in a sequential two-player inspection game, where players are assigned the role of either taxpayer or inspector. In this experiment, participants were assigned the role of taxpayer, while an AA was assigned the role of Inspector. The game consists of three rounds segments, each with 15 decisions, and an initial endowment of 100 tokens for each round segment. In each round participants can decide to either underreport their taxes or fully report their taxes. The corresponding payoff structure can be seen in Table 1. If the participant decides to underreport and the inspector decides to inspect, the taxpayer incurs a fine of 10 tokens. If the participant decides to underreport but does not get audited, they receive a payment of 5 tokens. To ensure symmetry in decisions, the same payoffs are used for the inspector, meaning a successful inspection results in a 5 token rewards, while an inspection on a full report leads to an inspection cost of 10 tokens. If both players do not underreport and inspect no balance change occurs. The inspector's decisions are based on pre-defined sequences derived from a previous inspection game (Rauhut, 2015), a methodology also employed by Schniter et al. (2020). Three decision sequences were extracted, with average inspection rates of 0.6, 0.53 and 0.4 across 15 rounds. The decision sequences are further used as control variables to ensure observed effects are not due to specific decision sequences of the AA while also providing more robust results along different inspection averages. The decision sequences of the inspectors were not known to the participants. Participants were informed of both players' decisions and their current balance after each decision round and of their final balance after each 15-round segment. The structured nature of this game, with the pre-defined decision sequences for the inspectors and the symmetric pay-off scheme, provides an ideal setup for studying strategic decision-making behaviour under controlled conditions.

Table 1: Payoff Structure for Taxpayer (Participant)



		Inspector	
		AUDIT	NO AUDIT
Taxpayer	UNDERREPORT	-10	5
	FULLY REPORT	0	0

3.2 Treatment

The study employs a within-subject design with three treatments, with each treatment being played for 15 rounds. In the first treatment, the human treatment, participants are told that they will play against a human inspector. In the second treatment, the simple bot treatment, they are informed that they will play against a simple algorithm. In the third treatment, the complex bot treatment, participants are informed that they would play against a complex AI that mimics human decisions. Importantly, the inspector plays out the same pre-defined decision sequences for all treatments, ensuring comparability between the three treatments. Participants are randomly allocated to one of six treatment sequences, covering all possible treatment orders (example sequence: 1. Human treatment, 2. Simple AA treatment, 3. Complex AA treatment). The information about the treatment is provided within the instructions at the start of the game in a separate paragraph to increase attention and focus on the treatment, as well as in a special page between the 15 round segments. The AA is further described as trying to gain as many tokens as possible, much like a human would, so that players felt that the inspector too had an incentive to detect tax evasion. The AA descriptions used for the different treatments can be found in Appendix 1. The use of descriptions to manipulate perceptions was also used by Lee et al. (2021) in their study, although they described the agent along agency and patiency dimensions. Nielsen et al. (2022) also used introductory statements to ensure players are aware that their counterpart is either human or computer to reinforce desired effects. This study uses AA terminology and capability descriptions to manipulate perception, borrowing from the findings of Langer et al. (2022) and their analysis of the impact of different AA terminology on perception. As an additional manipulation measure, the loading screen between decisions for the computer treatments differs from the human treatment, with the computer treatment showing a “waiting for computer” message, while the human treatment shows a “waiting for other player” message. Further, the participants are shown who they are playing against on top of each decision page. Upon the conclusion of the experiment, participants are informed about the manipulation of perception that took place in the experiment through a debrief page. By adopting a within-subject design for this experiment, we are not only able to examine initial treatment effects in the first set of 15 rounds, where participants were completely unaware of the existence of different agent types, but also observe how decision-making behavior evolves across different agent experiences in varying treatment sequence orders.

3.3 Survey Measures

Studies of human attitudes towards AI and machines have shown that socio-demographic factors, technical affinity, as well as knowledge of technology are critical factors that can influence perceptions on AA implementations. Examples include applications AI in healthcare (Fritsch et al. 2022), AI in decision-making (Kushwaha et al. 2022), general attitudes towards AI (Selwyn et al. 2020; Zhang and Dafoe 2019) and different forms of automated system applications (Langer 1992). Therefore, this experiment elicits such factors through several survey questionnaires. First, participants complete a 9-item questionnaire concerning their technical affinity. The 9 items are based upon the Affinity for Technology Interaction (ATI) scale by Franke et al. (2019), which are measured on a 6-point Likert scale from completely disagree to completely agree (see Appendix 2, Table 6). Second, they fill out a survey about their attitudes towards taxes which was built on segments from the Comprehensive Taxpayer Attitude Survey (2021)(see Appendix 3, Table 7). Attitudes toward taxation have been shown to be an influencing factor on decision-making in tax evasion, and therefore warrants inclusion as a control variable (Torgler, 2002; Wärneryd & Walerud, 1982). To ensure game understanding, treatment effectiveness, and gauge overall attitudes towards AAs, participants complete three additional surveys. After the first 15 decisions, participants are asked about their experience of playing against their specific treatment (see Appendix 4, Table 8 & 9). After the ATI and tax attitudes survey, a general survey is introduced where participants are asked about their level of understanding of the game, ensuring that the instructions and experimental procedure is clear (see Appendix 5, Table 10). Lastly, participants are asked about their attitudes towards AAs in general, as well as the use of AAs to control taxes and wider aspects of their life (see Appendix 6, Table 11). Collecting a broad spectrum of variables not only fortifies the robustness of subsequent analyses, but also enables deeper understanding into how decision-making could have been shaped by external factors.

3.4 Recruitment & Experiment

The platform used to program the experiment itself was O-Tree (Chen, Schonger, and Wickens 2016) using PyCharm, a Python IDE. For recruitment, the online web service Prolific was used. Prolific is an online research platform used to recruit study participants for research purposes, similar to services such as MTurk. Yet, studies have shown that Prolific provides more transparency for participants, offers better participant diversity and selection, as well as granting better functionality compared to MTurk and other services (Palan & Schitter, 2018; Peer et al., 2017). Previous research has also shown that online samples do not reduce data quality compared to traditional lab samples (Germine et al., 2012), and can show more diversity than traditional university student samples (Paolacci & Chandler, 2014). Nevertheless, researchers ought to be cautious in their employment of such tools, as they can come up with their own biases such as representing online populations. During the conduction of the experiment no problems were encountered, and the data was collected on a weekday afternoon where high participation rates are usually observed. The experiment was done in three waves, starting with a smaller wave to ensure no problems would be encountered. The payment for participants was based on a fixed fee (2£ for 20 minutes), plus a variable bonus based on performance. The median time for completion of the experiment was 19 minutes and 18 seconds, with a mean bonus payment of 2.56£, resulting in a mean hourly payment of 13.90£ across all experiment waves. Participants were informed that the bonus payment is calculated from one of the three round segments which will be randomly chosen as the payout relevant round segment at the end of the experiment. Participants were informed that their data would be kept anonymously, since the data was anonymized for the analysis.

4 Results

4.1 Descriptive

The experiment reached a final participant count of 300 individuals. We achieved overall high understanding of the experiment, tested with the post-experiment survey asking participants about their understanding of different components of the experiment (see Appendix 7, Figure 3). An average of 93.6% participants either ‘agreed strongly’ or ‘agreed’ to having understood the different components of the experiment and overall found no problems navigating the game. Concerning socio-demographic attributes, individual information about participants was acquired through the user data from Prolific. Individual variables were then grouped into larger groups to facilitate analysis, with NA designated to data that has expired. Table 2 gives an overview of the used categories. Regarding sex, a slight skew towards male participants was recorded, with 53.7% (n=161) of the participants identifying as male and 46% (n=138) female, with one person not wishing do disclose their sex. For other characteristics, the sample indicates that a majority of participants are under 30 with 73% (n=219), in paid work 57% (n=171), predominantly white 67% (n=201), and have non-English as their primary language 73% (221). The sample shows an overall high level of geographical diversity. Participants in our study were largely from South Africa, Poland, and Portugal, collectively accounting for more than 60% of the total sample. South Africa had the highest representation with 22.67% (n=68), followed by Poland at 19.33% (n=58), and Portugal at 19.00% (n=57). Italy and Greece accounted for a smaller portion of the sample, contributing 6.67% (n=20) and 4.67% (n=14) respectively. There were 37 Countries represented with the sample, although many of them were comprised of a single participant.

Table 2: Descriptive Distributions of Experiment Sample

Category Group	Category	Count	Percentage
Age Group	< 30	219	73.00%
	30-59	80	26.67%
	60 +	1	0.33%
Employment Grouped	in paid work	171	57.00%
	not in paid work	79	26.33%
	NA	50	16.67%
Ethnicity Grouped	White	201	67.00%
	Black	70	23.33%
	Other	28	9.33%
	NA	1	0.33%
Language Grouped	English	79	26.33%
	Non-English	221	73.67%
Sex	Male	138	46.00%
	Female	161	53.67%
	NA	1	0.33%
Top 5 Counties	South Africa	68	22.67%
	Poland	58	19.33%
	Portugal	57	19.00%
	Italy	20	6.67%
	Greece	14	4.67%

4.2 Decision Analysis

Answers from the initial treatment survey indicate significant differences between treatments, suggesting that the treatment manipulation was successful (see Appendix 8, Figure 4). Figure 1 shows the mean decision rates colored by treatment and grouped into three plots for the three decisions rounds, with 0 indicating no evasion and 1 representing a decision to evade. The line plot illustrates how decisions changed within the rounds as well as providing a comparison between the treatments and round groupings. The horizontal lines show the trend of the decisions, which are mostly stable in rounds 1-15. In rounds 16-30, we see decisions in the simple treatment declining throughout the round segment, while participants in the human treatment were more likely to evade in the later rounds. In the last round segment, rounds 31-45, participants started off with higher rates tax evasion, but in all treatments this behavior declined over the course of the round segment.

Figure 1: Line plot of Average Decision to Evade by Treatment and

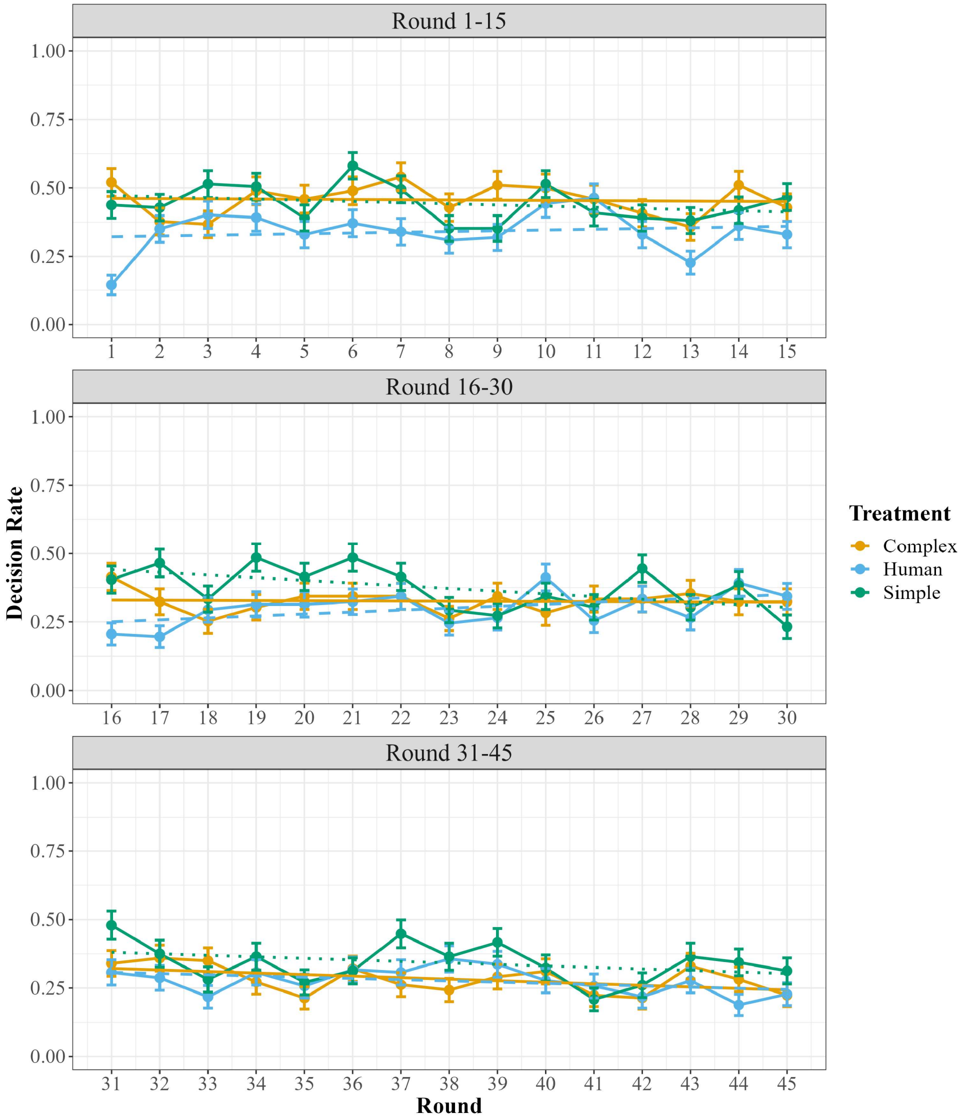
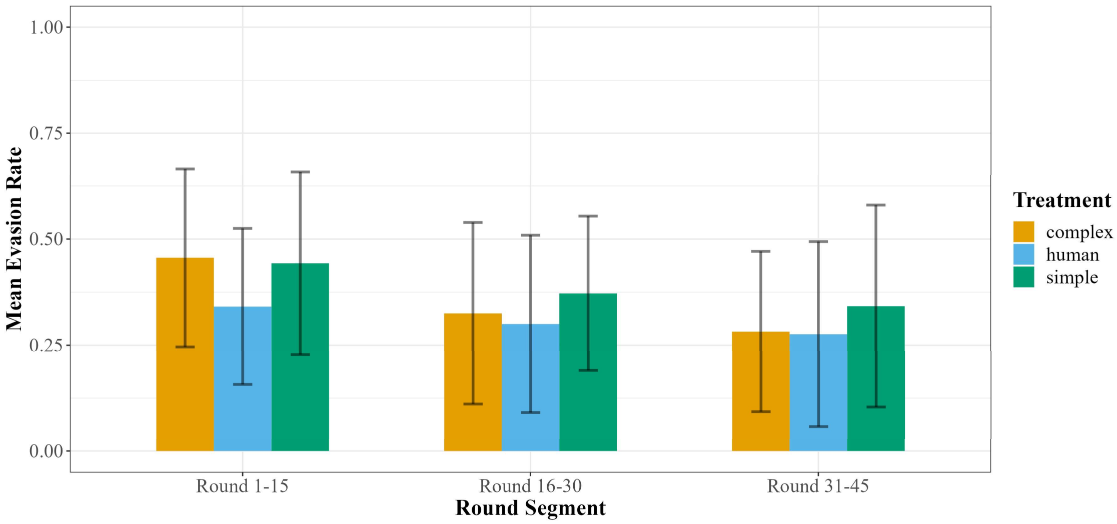


Table 3 and Figure 2 present the mean evasion rates, count, and standard deviation across treatments and round groups. In the first round segment (rounds 1-15), both complex (mean 0.456, SD 0.21) and simple (mean 0.443, SD 0.216) treatments exhibit higher evasion rates than the human treatment (mean 0.341, SD 0.184). In the second round segment (rounds 16-30), the complex treatment sees a large drop in evasion rates (mean 0.325, SD 0.214), aligning more closely with the human treatment (mean 0.3, SD 0.209), while the simple treatment also decreases, albeit less pronounced (mean 0.372, SD 0.182). By the final round segment (rounds 31-45), the complex treatment (mean 0.282, SD 0.189) nearly matches the human treatment (mean 0.276, SD 0.218), while the simple treatment continues to show higher evasion rates (0.342, SD 0.238), despite a decline. Notably, the high standard deviations suggest diverse strategies among participants, persisting throughout the whole experiment and across treatments. However, the complex treatment sees a slight drop in SD from 0.214 in Round 16-30 to 0.189 in Round 31-45. The human treatment's SD consistently increases across rounds, from 0.184 in round 1-15, to 0.209 in round 16-30 and 0.218 in rounds 31-45. The simple treatment sees a drop in SD between rounds 1-15 (0.216) and 16-30 (0.182), but an increase for rounds 31-45 (0.238).

Table 3: Mean Evade Decisions grouped by Treatment and Round Segment

	Complex			Human			Simple		
	Mean Evasion	SD	Count	Mean Evasion	SD	Count	Mean Evasion	SD	Count
Round 1-15	0.456	0.21	98	0.341	0.184	97	0.443	0.216	105
Round 16-30	0.325	0.214	99	0.3	0.209	102	0.372	0.182	99
Round 31-45	0.282	0.189	103	0.276	0.218	101	0.342	0.238	96

Figure 2: Boxplot of Mean Evade Decision Rates by Treatment and Round



To assess the statistical significance of treatment effects, a regression analysis was conducted. Initially, a linear regression analysis was performed, owing to its capacity to incorporate multiple independent variables and identify discrepancies across treatment means and round segments. Table 4 presents the results of the linear regression model with mean evasion rates across rounds, and three models estimated for each round segment. In the first exposure to treatment (rounds 1-15), a significant difference emerges in evade decisions between participants in the human treatment and the reference category of the complex AA treatment, evidenced by a coefficient of -0.111 (p-value < 0.01). This suggests that participants that perceived to be playing against a human saw a substantial decrease in mean tax evasion of 0.111 compared to the reference category of the complex treatment. This finding supports the first hypothesis, with participants more likely to evade taxes in the AA treatments, albeit only for the initial 15 rounds. In round 16-30, there is no significant effect of the treatments for both the simple and human treatment. Yet, in rounds 31-45 we see a significant positive effect between the simple and complex treatment with a coefficient of 0.064 (p-value <0.05). This indicates that participants were more likely to evade taxes against the simple computer in the last group rounds of 31-45 compared to the complex treatment. Thus, the second hypothesis is also supported by these findings, wherein a significant difference in evasion behavior between perceived complex and simple AAs emerges, although this is only evident in rounds 31-45. Interestingly, neither socio-demographic factors, technical affinity nor tax attitudes significantly influenced these evade decisions except for ethnicity in round 31-45 with a coefficient of 0.104 (p-value < 0.1), where people assigning themselves to black ethnicity are more likely to evade taxes. An examination of the distribution of the Tax and ATI measures reveals that both variables predominantly adhere to a normal distribution (see Appendix 9, Figures 5 & 6). This suggests that they are unlikely to introduce issues in the statistical analyses due to skewness or outliers.

Table 4: Linear Regression Model of Mean Evasion Rates

	Dependent variable: Mean Evade		
	Round 1-15 (1)	Round 16-30 (2)	Round 31-45 (3)
Human Treatment (Ref: Complex Treatment)	-0.111*** (0.030)	-0.027 (0.029)	-0.002 (0.031)
Simple Treatment (Ref: Complex Treatment)	-0.007 (0.029)	0.045 (0.029)	0.064** (0.031)
Male (Ref: Female)	-0.010 (0.026)	0.011 (0.026)	-0.018 (0.028)
Age: 30-39 (Ref: <30)	-0.019 (0.028)	-0.012 (0.027)	-0.022 (0.030)
Ag: 60+ (Ref: <30)	-0.177 (0.207)	-0.158 (0.203)	0.142 (0.218)
Ethnicity: Black (Ref: White)	0.074 (0.058)	0.083 (0.057)	0.104* (0.061)
Ethnicity: Other (Ref: White)	-0.027 (0.043)	-0.015 (0.042)	0.031 (0.045)
Non-English Main Language (Ref: English)	0.024 (0.054)	-0.008 (0.053)	0.075 (0.057)
Technological Affinity	-0.002 (0.018)	-0.006 (0.018)	-0.026 (0.019)
Attitude on Tax	-0.001 (0.017)	-0.00002 (0.017)	-0.005 (0.018)
Constant	0.445*** (0.110)	0.339*** (0.109)	0.340*** (0.115)

Observations	298	298	298
R^2	0.082	0.061	0.042
Adjusted R^2	0.050	0.029	0.009
Residual Std. Error (df = 287)	0.205	0.201	0.216
F Statistic (df = 10; 287)	2.550^{***}	1.872^{**}	1.264

*Note: Significant noted as *p<0.1; **p<0.05; ***p<0.01. Linear regression results with dependent variable mean evade decisions for each round segment. Independent variables include factor variable for treatment condition ('human', 'simple', with complex as the reference category), categorical variables sex ('Male' with 'Female' as the reference category), age group ('30-59', '60 and above' with 'Under 30' as the reference category), ethnicity group ('Black', 'Other' with the reference category "White"), language group ('Non-English' with 'English' as the reference category), and continuous variables 'ATI Scale and 'Attitudes towards Taxes'.*

Given that a linear regression model on mean decision rates might not fully leverage the repeated binary decision design of this experiment, a mixed-effects logistic regression model was estimated, as shown in Table 5. This method aptly handles binary decision data, such as the decision to evade or not to evade taxes, while also accommodating for random effects between participants and treatment sequences. The model was fitted using data from all rounds, which were further divided into the three round segments: 1-15, 16-30, and 31-45. The results show that the human treatment had a statistically significant negative effect on evasion during rounds 1-15 (coefficient = -0.508, p < 0.01) and across all rounds (coefficient = -0.237, p < 0.01) compared to the complex treatment reference category. These findings reinforce the rejection of the null hypothesis associated with hypothesis 1, suggesting a clear disparity in evasion behavior between the human treatment and the AA treatments, with the latter demonstrating overall higher evasion rates. In contrast, the simple treatment had a significant positive effect on evasion during rounds 16-30 (coefficient = 0.245, p < 0.1) and rounds 31-45 (coefficient = 0.323, p < 0.1). This might be reflective of the decrease of the evasion rates in the complex treatment found in the later round segments while the evade rates in the simple treatment stayed more consistent. Over all rounds, the simple treatment showed a significant positive effect on evasion (coefficient = -0.161, p < 0.01). This further solidifies the support for hypothesis 2, with the simple AA treatment showing higher evasion rates compared to the AA complex treatment.

To enhance the robustness of the analysis, several independent variables outside of the treatment variables were incorporated. The decision sequence, denoting the inspection rate of the automated inspector, showed a significant effect across rounds and sequence variations. The decision sequence with a 0.43 mean inspection rate (43% inspection out of 15 rounds) was found to have a significant positive effect on evading during rounds 16-30 (coefficient 0.431, p < 0.01), rounds 31-45 (coefficient = 0.963, p < 0.01), as well as all rounds showing a significant effect (coefficient = 0.459, p < 0.01) compared to the reference category of 0.5 mean inspection rate. Similar effects can be seen when considering the decision sequence with a 0.6 mean inspection rate, which also showed positive significant in rounds 16-30 (coefficient 0.416, p < 0.01) and rounds 31-45 (coefficient 0.438, p < 0.01), as well as overall rounds (coefficient 0.244, p < 0.01). Therefore, decision sequence showed a positive effect on participants decision to evade taxes, yet this can be seen across different mean inspection rates. On the other hand, the round variable shows a significant negative effect in rounds 31-45 (coefficient = -0.030, p < 0.01) and across all rounds (coefficient = -0.012, p < 0.01). This indicates that with each round within the 15 round segments the log-odds of evading decreases by -0.012. An independent variable was added to indicate if the participant was caught evading by the inspector in the previous round. This variable showed a substantial and significant effect in rounds 1-15 (coefficient = 0.706, p < 0.01), which prevailed when considering all rounds (coefficient = 0.416, p < 0.01) Lastly, male gender did not have an effect on evasion across all rounds (coefficient = -0.068, p > 0.1). For simplicity and clarity in the analysis and to facilitate better understanding and interpretation of the results, other control variables which were analyzed are excluded in the final analysis (see Appendix 10, Table 12 for the full analysis). While the control variables had slight impacts, the significance of the main effects persisted throughout all variations of the model.

Table 5: Mixed Effects Logistic Regression Model of Evade Decisions

	Dependent variable: Evade				
	R: 1-15	R: 16-30	R: 31-45	All Rounds	All Rounds sex
	(1)	(2)	(3)	(4)	(5)
Human Treatment (*Ref:*	-0.508***	-0.143	-0.080	-0.237***	-0.232***
Complex Treatment)	(0.130)	(0.146)	(0.162)	(0.048)	(0.048)
Simple Treatment (Ref:	-0.055	0.245*	0.323**	0.161***	0.167***
Complex Treatment)	(0.126)	(0.145)	(0.162)	(0.047)	(0.047)
Decision Sequence 0.4	0.144	0.431***	0.963***	0.459***	0.458***
(Ref: 0.5)	(0.129)	(0.147)	(0.165)	(0.122)	(0.122)
Decision Sequence 0.6	-0.040	0.416***	0.438***	0.244**	0.246**
(Ref: 0.5)	(0.126)	(0.144)	(0.163)	(0.120)	(0.120)
Round	-0.006	-0.005	-0.030***	-0.012***	-0.012***
	(0.008)	(0.008)	(0.008)	(0.004)	(0.004)
Caught Previous Round	0.706***	0.095	0.185*	0.416***	0.418***
	(0.081)	(0.089)	(0.099)	(0.049)	(0.049)
Male (Ref: Female)					-0.068
					(0.099)
Constant	-0.347***	-1.101***	-1.369***	-0.912***	-0.880***
	(0.132)	(0.148)	(0.163)	(0.099)	(0.113)
Observations	4,500	4,500	4,500	13,500	13,455
Log Likelihood	-2,881.830	-2,718.994	-2,524.287	-8,116.207	-8,088.112
Akaike Inf. Crit.	5,779.660	5,453.988	5,064.573	16,250.410	16,196.220
Bayesian Inf. Crit.	5,830.955	5,505.283	5,115.868	16,318.010	16,271.300

*Note: Significant noted as *p<0.1; **p<0.05; ***p<0.01. Mixed effects logistic Regression Model. Treatment reference category is complex treatment for all models. Decision Sequence is depicting the three difference inspection sequences with different mean inspection rates used for the inspection algorithm. Rounds are round numbers 1-15 for each decision made within the round groups. Model 1-3 have random intercept for participant id, showing treatment effects for that specific round group. Model 4-6 span all round groups and have random* *intercept for both Participant ID as well as Participant and treatment sequence combinations (6 in total) to reflect within subject design considerations. Difference in n for model 5 due to 1 missing data point in gender.*

4.3 Further Decision Analysis

While both hypotheses were substantiated by the main effects of the treatments, a detailed exploration of the significant confounding factors, identified within the regression analyses, could provide further valuable insights. Firstly, the variable 'round' in table 3, which represents changes within decision-making in each round, exhibits an overall significant effect. To understand how rounds affect participants' decision in more detail, the individual round segments have been aggregated to examine overall trends of evasion decisions over the 15 rounds for each treatment (See Appendix 11, Figure 7 & 8). Discernable negative trends in both complex and simple treatments reveal participants becoming less likely to evade taxes as the rounds progress, with the simple treatment showing a more pronounced decline. The human treatment on the other hand shows a slight incline, with participants on average being more likely to evade taxes as the rounds progress. These trends were further examined within a mixed effects logistic regression (Appendix 11, Table 13), where the negative effect of round number on participants in the simple treatment showed a significant effect, reducing log odds of evasion with each subsequent round (coefficient = 0.032, $p < 0.01$). Furthermore, the interaction between treatment and round number was assessed. We find a significant interaction effect for both simple and human treatment with round numbers compared to the complex treatment. The positive coefficient for the interaction term between human treatment and round number (coefficient = 0.021, $p < 0.1$) indicates that for participants in the 'human' treatment group, the likelihood of evasion decreases less with each additional round relative to the complex treatment group. Conversely, the negative coefficient for the interaction simple treatment and round number (coefficient = -0.020, $p < 0.1$) suggests a more pronounced decrease in evasion likelihood per round in the 'simple' treatment group compared to the complex treatment. These findings suggest that the influence of round number on evasion behavior may depend on the specific treatment, as well as the treatment effects depending on the round number, although these interaction effects were restricted in their statistical significance with p-values between 0.1 and 0.05.

The second additional analysis was done with data gathered through the additional surveys deployed within the experiment. Firstly, after the first 15 decisions and the initial exposure to the treatment, participants were queried about their experience with the treatment-dependent inspector (see Appendix 12, Figure 9). A Wilcoxon rank sum test on the survey data suggests that AAs were perceived as more complex in the complex treatment ($p < 0.01$) and simpler in the simple treatment ($p < 0.01$), affirming the treatment effect. Other variables indicate that some participants evaluated the AAs as more strategic and human-like in their decisions, while others disagreed with such sentiments. In order to ascertain if such experiences influence the decisions in the first 15 rounds, a mixed effects logistic regression model was run for the corresponding scores on the Likert scale and decision to evade taxes as the dependent variable (see Appendix 12, Table 14). Yet, none of the variables were found to be significantly affecting evasion decisions. Secondly, at the end of the experiment participants were asked about their general sentiments towards AAs and the use of AAs as inspectors (see Appendix 13, Figure 10). Using these variables, a linear regression was run using the mean decisions to evade taxes as a dependent variable as well as a mean survey score and the individual answers as independent variables (see Appendix 13, Table 15). General considerations of fairness and objectiveness of AAs did not have a significant effect on mean evasion decisions, but answers to the question “I would support the implementation of AAs to control wider areas of my life" showed a significant effect on participants decisions over all treatments (coefficient = 0.036, $p < 0.05$). More precisely, participants in the human treatment, who showed higher agreement with that sentiment were more likely to evade taxes (coefficient = 0.049, $p < 0.01$), with participants in the complex treatment also being more likely to evade taxes (coefficient = 0.037, $p > 0.05$), and no effect identified on participants in the human treatment.

5 Discussion

The objective of this study was to augment the existing literature on human-computer interaction by introducing an inspection game that included AAs as inspectors. Tax evasion was used as a framework to immerse participants in a specific norm-deviating context, building upon the foundation which was set by previous works in this field. Both tested hypotheses were supported by the results: participants were more likely to evade taxes when dealing with AAs compared to a human (Hypothesis 1), and they were more likely to evade taxes when interacting with an agent described to be simpler compared to a complex one (Hypothesis 2). These findings align with previous research (Maréchal et al., 2020; De Melo et al., 2016), suggesting that individuals are more likely to exploit machines in strategic exchanges. The fact that the complex treatment converged with the human treatment over the course of the experiment indicates the potential of participants evaluating humans and complex agents similarly. This in turn could indicate the appliance of similar expectations and norms within a strategic exchange once familiarity with different agents is reached (Reeves and Nass 1996). Additionally, this aligns with the concept of mindlessness (Langer et al. 2022), suggesting that participants may enter a mental state in which they automatically apply social scripts during interactions with complex AAs. It could also confirm that participants estimate the perceived capabilities of AAs described as complex more highly, showing higher trust in their capabilities (Robinette et al. 2016; Salem et al. 2015), and therefore see higher risk associated with evading taxes.

From an experimental perspective, it’s important to remark on the internal variables which also influenced decisions. We saw that the decision sequences used by the automated bot inspectors had a significant influence on the decisions of participants. Combined with the finding that getting caught has a positive influence on evasion the following round, it can be presumed that decision sequences with higher inspection rates lead to more catches of evasions. Participants seemed to be more likely to try to evade taxes after being caught, hinting at the opposite of a deterrence effect. Incorporating round numbers into the regression analysis revealed a significant negative effect, demonstrating a decreasing tendency in evasion behavior as participants progressed through consecutive rounds. However, this effect was found to vary between different round segments and treatments, with significant interaction effects noted between rounds and treatments. Specifically, the human treatment group demonstrated a slower decrease in evasion likelihood with each additional round, while the simple treatment group displayed a more pronounced reduction in evasion likelihood per round, both relative to the complex treatment group. Lastly, the inclusion of several independent variables ensured the robustness of results. Most socio-demographic factors did not show any significant influence except ethnicity which showed slightly higher mean evasion rates. Measures for both technical affinity and overall tax sentiments were also employed in order to ensure that such factors were accounted for. Against expectations, both technical affinity and tax sentiments did not have any significant influence on tax evasion decisions. While the values seem to be distributed normally, there is still a possibility that biases are introduced by the nature of the online sample. These last findings go against what has been found in ATI literature and Tax literature, where such sentiments were deemed to have an influence on decisions and interactions (Franke et al., 2019; Torgler, 2002; Wärneryd & Walerud, 1982).

The post-experiment survey supports the efficacy of the methodology used, where participants did recognize playing against an AA, and perceived its complexity or simplicity dependent on the treatment conditions. Specifically, it adheres to the notions of Nass and Moon (2000) and Nass et al. (1994), supposing that humans can perceive machines to emulate human behavior and strategy, which in turn influences their behavior towards them, as well as their strategic outlook. An intriguing result from the post-experiment survey is the fact that participants rated the computer as a more objective and slightly fairer inspector, but nevertheless preferred playing against a human within the experiment. The importance here is that humans have shown a general preference towards interacting with humans over machines in social decision-making scenarios, which has previously been identified in other studies (Gallagher et al. 2002; McCabe et al. 2001). Additionally, humans evaluate fairness between AAs and fellow humans differently (Wang et al., 2020), which can lead to higher perceptions of fairness, but can also ultimately lead to preferring a human inspector (De Melo et al., 2016). Such inherent characteristics might be reflected within the results of the experiment. Linear Regression results have also shown that the support towards AA in broader aspects of life has a significant effect of evasion rates when playing against humans, where higher evasion rates could be seen in people disagreeing with such a notion. The interplay between not wanting AA control to be implemented more broadly and exhibiting higher norm-deviating behavior against AA could be linked to general perceptions of AAs capabilities and subsequent risk estimations.

In general, findings within this experiment have both theoretical as well as practical implications. From a theoretical standpoint, the study allows an extension of previous HCI experiments, where the duality of human and non-human agents is tested in the new context of a non-cooperative game setting. This allows the study to both identify differences between decision-making in the different interaction settings, as well as developing a deeper understanding of how humans react to automated supervision, which will increasingly become relevant in today's technologically controlled world. From a practical standpoint, findings can inform both public and private institutions in their application of AAs as controlling mechanisms. If the AAs do not deter norm deviating or deviant behavior, institutions might consider withholding wider implementation of such technologies, or at least weight the increase in norm deviating behavior against cost-saving benefits. Further, simply describing an agent to be complex can reduce norm-deviating behavior, and potentially align performance closer to a human inspector. Future research might expand on these findings in other contexts, while also addressing personal preferences and strategic estimations of human and AAs. It is important to keep in mind how the public, and specifically the people being supervised, feel, and react to such supervision. As has been cautiously illustrated in the post-experiment survey, while people might rate computers as being more objective and fairer, they might still prefer human supervision. An informed discussion should take place where risks, benefits and perceptions of the affected persons are considered critically in order to pave the way for sustainable development and implementation of such technology.

6 Limitations

Firstly, the study population was recruited on an online platform, which has a higher likelihood of consisting of participants that have higher technological expertise and positive viewpoints of technology, therefore potentially reducing the generalizability of results. The ATI scale was used to measure this phenomenon, but it does not completely eradicate effects perceived through this imbalance. Secondly, studies in this area have used a variety of different denotations to label AAs, from computers to algorithms, all the way to artificial intelligence. As Langer et al. (2022) have shown, terminology does affect perceptions, and therefore the terminology should be employed with care and critical reflection. While this paper has taken such notions into consideration, it is important to recognize the possibility that the terminology used within the instructions of the experiment can lead to adverse effects on participants, with different participants having different perceptions of specific denotations. This is also true for the tax framework, where different studies employ different terminologies, which in turn can influence the strategic decision-making. Thirdly, the algorithm used in this experiment is based upon pre-defined sequences taken from a previous inspection game experiment. While it is unlikely that the human players noticed the pre-defined nature of the decisions, it can nevertheless undermine the strategic nature of the inspection game, where decisions are based upon previous decisions of your opponent. Employing AAs that play on a defined strategy but react to decisions by the participants might overcome such limitations. Finally, the treatments in this experiment are grounded in the manipulation of participants' perceptions. Thus, the effects observed are constrained to how the agents were perceived, rather than the experience of actually interacting with these agents. This presents a limitation to the external validity of the study, as findings may not be directly applicable to scenarios involving interactions with different, actual agents. In future studies, it could be beneficial to introduce a real human and an actual complex AA into the experiment. The human role could be played by human participants, while a learning algorithm similar to the one used by Ishowo-Oloko et al. (2019) could be used for the complex AA. Despite these limitations impacting the interpretation of the findings, the study provides compelling evidence of differences in decision-making when perceptions alone are manipulated.

Ethics Statement

The study involving human participants were reviewed and approved by University of Zurich's Ethics Committee of the Faculty of Arts and Social Sciences, #23.04.22. The participants provided their informed consent to participate in this study.

Acknowledgments

Thanks to Prof. Dr. Heiko Rauhut for making this study possible. Thanks to Vincent Oberhauser and Dr. Fabian Winter for their invaluable help and insights which helped shape this study and turn it from an abstract idea into reality. Thanks also to Strahinja Popovic, Diego Strassmann Rocha and Prof. Dr. Hagen Worch for their final comments and reviews. Finally, thank you to all my other peers who have given me continuous feedback along the way.

Bibliography

Alm, J. (2012). Measuring, explaining, and controlling tax evasion: Lessons from theory, experiments, and field studies. *International Tax and Public Finance*, *19*(1), 54–77. https://doi.org/10.1007/s10797-011-9171-2

Andreozzi, L. (2004). Rewarding policemen increases crime. Another surprising result from the inspection game. *Public Choice*, *121*(1–2), 69–82. https://doi.org/10.1007/s11127-004-6166-x

Baldry, J. C. (1986). Tax evasion is not a gamble: A report on two experiments. *Economics Letters*, *22*(4), 333–335. https://doi.org/10.1016/0165-1765(86)90092-3

Beck, P., Davis, J. S., & Jung, W. O. (1991). Experimental Evidence on Taxpayer Reporting under Uncertainty. *Accounting Review*, *66*(3), 535–588.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, *76*(2), 169–217. https://doi.org/10.1086/259394

Bianco, W. T., Ordeshook, P. C., & Tsebelis, G. (1990). Crime and Punishment: Are One-Shot, Two-Person Games Enough? *American Political Science Review*, *84*(2), 569–586. https://doi.org/10.2307/1963536

Blumenthal, M., Christian, C., & Slemrod, J. (2001). Do Normative Appeals Affect Tax Compliance? Evidence from a Controlled Experiment in Minnesota. *National Tax Journal*, *54*(1), 125–138. https://doi.org/2021031208092500661

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Comprehensive Taxpayer Attitude Survey (CTAS) 2021 (Publication 5296 (Rev. 4–2022) Catalog Number 71353Y). (n.d.). US Department of the Treasury Internal Revenue Service. https://www.irs.gov/pub/irs-pdf/p5296.pdf

Coricelli, G., Joffily, M., Montmarquette, C., & Villeval, M. C. (2010). Cheating, emotions, and rationality: An experiment on tax evasion. *Experimental Economics*, *13*(2), 226–247. https://doi.org/10.1007/s10683-010-9237-5

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-017-02597-8

Daylamani-Zad, D., & Angelides, M. C. (2021). Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games. *Ieee Transactions on Games*, *13*(3), 229–238. https://doi.org/10.1109/TG.2020.2989636

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people’s minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*, 73–88. https://doi.org/10.1037/a0034251

De Melo, C., Marsella, S., & Gratch, J. (2016). People Do Not Feel Guilty About Exploiting Machines. *Acm Transactions on Computer-Human Interaction*, *23*(2), 8. https://doi.org/10.1145/2890495

Dresher, M. (1962). *A Sampling Inspection Problem in Arms Control Agreements: A Game-Theoretic Analysis*. RAND Corporation. https://www.rand.org/pubs/research_memoranda/RM2972.html

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *The British Journal of Social Psychology*, *51*(4), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the Intentional Stance in a Competitive Game. *NeuroImage*, *16*(3, Part A), 814–821. https://doi.org/10.1006/nimg.2002.1117

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), Article 1. https://doi.org/10.1038/s41591-018-0307-0

Kaber, D. B. (2018). A conceptual framework of autonomous and automated agents. *Theoretical Issues in Ergonomics Science*, *19*(4), 406–430. https://doi.org/10.1080/1463922X.2017.1363314

Karpus, J., Krueger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *Iscience*, *24*(6), 102679. https://doi.org/10.1016/j.isci.2021.102679

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner’s dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology*, *70*, 47–65. https://doi.org/10.1037/0022-3514.70.1.47

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLOS ONE*, *3*(7), e2597. https://doi.org/10.1371/journal.pone.0002597

Langer, E. J. (1992). Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition*, *1*(3), 289–305. https://doi.org/10.1016/1053-8100(92)90066-J

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). ‘Look! It’s a Computer Program! It’s an Algorithm! It’s AI!’: Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems? *CHI Conference on Human Factors in Computing Systems*, 1–28. https://doi.org/10.1145/3491102.3517527

Lee, M., Lucas, G., & Gratch, J. (2021). Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. *Journal on Multimodal User Interfaces*, *15*(2), 201–214. https://doi.org/10.1007/s12193-020-00356-6

Lefebvre, M., Pestieau, P., Riedl, A., & Villeval, M. C. (2015). Tax evasion and social information: An experiment in Belgium, France, and the Netherlands. *International Tax and Public Finance*, *22*(3), 401–425. https://doi.org/10.1007/s10797-014-9318-z

Leslie, D. (2020). *Understanding bias in facial recognition technologies*. https://doi.org/10.5281/zenodo.4050457

Lohr, S. (2022). Facial Recognition Is Accurate, If You're a White Guy *. In *Ethics of Data and Analytics*. Auerbach Publications.

Maréchal, M., Cohn, A., & Gesche, T. (2020). Honesty in the digital age. *Working Paper Series / Department of Economics*, *280*, Article 280. https://doi.org/10.5167/uzh-149945

Mascagni, G. (2018). From the Lab to the Field: A Review of Tax Experiments. *Journal of Economic Surveys*, *32*(2), 273–301. https://doi.org/10.1111/joes.12201

Maschler, M. (1966). A price leadership method for solving the inspector's non-constant-sum game. *Naval Research Logistics Quarterly*, *13*(1), 11–33. https://doi.org/10.1002/nav.3800130103

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835. https://doi.org/10.1073/pnas.211415698

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678. https://doi.org/10.1006/ijhc.1996.0073

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Moon, Y., & Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology*, *27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703

Nielsen, Y. A., Thielmann, I., Zettler, I., & Pfattheicher, S. (2022). Sharing Money With Humans Versus Computers: On the Role of Honesty-Humility and (Non-)Social Preferences. *Social Psychological and Personality Science*, *13*(6), 1058–1068. https://doi.org/10.1177/19485506211055622

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2022, November 30). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. https://openai.com/blog/chatgpt/

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Radu, S. (2015). Multi-Issue Automated Negotiation with Different Strategies for a Car Dealer Business Scenario. In I. Dumitrache, A. M. Florea, F. Pop, & A. Dumitrascu (Eds.), *2015 20th International Conference on Control Systems and Computer Science* (pp. 351–356). Ieee. https://doi.org/10.1109/CSCS.2015.53

Rauhut, H. (2015). Stronger inspection incentives, less crime? Further experimental evidence on inspection games. *Rationality and Society*, *27*(4), 414–454. https://doi.org/10.1177/1043463115576140

Rauhut, H., & Jud, S. (2014). Avoiding Detection or Reciprocating Norm Violations? An Experimental Comparison of Self- and Other-Regarding Mechanisms for Norm Adherence. *Soziale Welt-Zeitschrift Fur Sozialwissenschaftliche Forschung Und Praxis*, *65*(2), 153–183. https://doi.org/10.5771/0038-6073-2014-2-153

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios: 11th Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2016. *HRI 2016 - 11th ACM/IEEE International Conference on Human Robot Interaction*, 101–108. https://doi.org/10.1109/HRI.2016.7451740

Roose, K. (2023, January 12). Don’t Ban ChatGPT in Schools. Teach With It. *The New York Times*. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 141–148. https://doi.org/10.1145/2696454.2696497

Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, *78*, 102253. https://doi.org/10.1016/j.joep.2020.102253

Sestino, A., Peluso, A. M., Amatulli, C., & Guido, G. (2022). Let me drive you! The effect of change seeking and behavioral control in the Artificial Intelligence-based self-driving cars. *Technology in Society*, *70*, 102017. https://doi.org/10.1016/j.techsoc.2022.102017

Spicer, M. W., & Thomas, J. E. (1982). Audit probabilities and the tax evasion decision: An experimental approach. *Journal of Economic Psychology*, *2*(3), 241–245. https://doi.org/10.1016/0167-4870(82)90006-X

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*, *613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z

Torgler, B. (2002). Speaking to Theorists and Searching for Facts: Tax Morale and Tax Compliance in Experiments. *Journal of Economic Surveys*, *16*(5), 657–683. https://doi.org/10.1111/1467-6419.00185

Torgler, B. (2007). *Tax Compliance and Tax Morale* [Books]. Edward Elgar Publishing. https://econpapers.repec.org/bookchap/elgeebook/4096.htm

Tsebelis, G. (1989). The Abuse of Probability in Political Analysis: The Robinson Crusoe Fallacy. *American Political Science Review*, *83*(1), 77–91. https://doi.org/10.2307/1956435

Tsebelis, G. (1990). Penalty has no Impact on Crime: A Game-Theoretic Analysis. *Rationality and Society*, *2*(3), 255–286. https://doi.org/10.1177/1043463190002003002

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376813

Wärneryd, K.-E., & Walerud, B. (1982). Taxes and economic behavior: Some interview data on tax evasion in Sweden. *Journal of Economic Psychology*, *2*(3), 187–211. https://doi.org/10.1016/0167-4870(82)90003-4

Webley, P., & Halstead, S. (1986). Tax Evasion on the Micro: Significant Simulations or Expedient Experiments? *Journal of Interdisciplinary Economics*, *1*(2), 87–100. https://doi.org/10.1177/02601079X8600100204

Weiss, M., Rodrigues, J., Paelecke, M., & Hewig, J. (2020). We, Them, and It: Dictator Game Offers Depend on Hierarchical Social Status, Artificial Intelligence, and Social Dominance. *Frontiers in Psychology*, *11*, 541756. https://doi.org/10.3389/fpsyg.2020.541756

Appendix

Appendix 1 – Treatment Descriptions of Automated Agents in the Experiment Introduction

(Complex) Automated Agent

The automated agent used within this experiment is a complex artificial intelligence (AI) agent that makes its own choices. The AI learned from previous experimental data and will emulate human decision-making behavior. The AI will try to increase its own tokens.

(Simple) Automated Agent

The automated agent used within this experiment is a simple algorithm with a basic predefined decision matrix. The algorithm will try to increase its own tokens.

(Human) Inspector

The Inspector will be played by another participant of this study who is tasked and paid to detect as many false declarations as possible to reduce tax fraud.

Appendix 2 - Affinity for Technology Interaction (ATI) scale by Franke et al. (2019)

Table 6: ATI Survey

In the following questionnaire, you will be asked about your interaction with technical systems. The term "technical systems" refers to apps and other software applications, as well as entire digital devices (e.g., mobile phone, computer, TV, car navigation)

Please indicate the degree to which you agree/disagree with the following statements	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
1 I like to occupy myself in greater detail with technical systems.	○	○	○	○	○	○
2 I like testing the functions of new technical systems.	○	○	○	○	○	○
3 I predominantly deal with technical systems because I have to.	○	○	○	○	○	○
4 When I have a new technical system in front of me, I try it out intensively.	○	○	○	○	○	○
5 I enjoy spending time becoming acquainted with a new technical system.	○	○	○	○	○	○
6 It is enough for me that a technical system works; I don't care how or why.	○	○	○	○	○	○
7 I try to understand how a technical system exactly works.	○	○	○	○	○	○
8 It is enough for me to know the basic functions of a technical system.	○	○	○	○	○	○
9 I try to make full use of the capabilities of a technical system.	○	○	○	○	○	○

Appendix 3 – Tax Attitude Survey

Table 7: Tax Attitude Survey

	Please indicate the degree to which you agree/disagree with the following statements	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
1	I support taxes in general	○	○	○	○	○	○
2	I believe the overall concept of taxing to be fair	○	○	○	○	○	○
3	I believe my current tax system to be fair	○	○	○	○	○	○
4	I report my taxes truthfully because I fear an audit and fines	○	○	○	○	○	○
5	I report my taxes truthfully because I beleive it's the right thing to do	○	○	○	○	○	○
6	I believe taxes help the government take care of citizens and national interests	○	○	○	○	○	○
7	I am suspicious on how the government uses tax money	○	○	○	○	○	○
8	I believe that people close to me are reporting and paying their taxes honestly	○	○	○	○	○	○
9	I believe that people overall report and pay their taxes honestly	○	○	○	○	○	○
10	I believe underreporting taxes to be a serious crime	○	○	○	○	○	○
11	I believe measures should be taken to reduce tax evasion	○	○	○	○	○	○

Appendix 4– Post Decision Survey (First 15 Rounds)

Human Treatment

Table 8: Post Decision Quiz (Human Treatment)

	Please indicate the degree to which you agree/disagree with the following statements	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
1	I believe the human Inspector made strategic decisions	○	○	○	○	○	○
2	I believe the human Inspector reacted to my decisions	○	○	○	○	○	○
3	I believe the human Inspector tried to increase their own payout	○	○	○	○	○	○
4	I believe the human Inspector to be fair	○	○	○	○	○	○

Simple & Complex Treatment

Table 9: Post Decision Quiz (Automated Agent Treatments)

	Please indicate the degree to which you agree/disagree with the following statements	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
1	I was aware of playing against an automated agents and not a human	○	○	○	○	○	○
2	I perceived the automated agents as being simple	○	○	○	○	○	○

3	I perceived the automated agents as being complex	○	○	○	○	○	○
4	I believe the automated agents made strategic decisions	○	○	○	○	○	○
5	I perceived the automated agents to emulate human behaviour in it's decision making	○	○	○	○	○	○
6	I found it hard to outsmart the automated agent	○	○	○	○	○	○

Appendix 5 – Post-Experiment Game Understanding Survey

Table 10: Post-Experiment Game Understanding Survey

	Please indicate the degree to which you agree/disagree with the following statements	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
1	I understood the outline of the game	○	○	○	○	○	○
2	I understood how the roles were assigned	○	○	○	○	○	○
3	I understood the decision making process	○	○	○	○	○	○
4	I understood the payoff structure	○	○	○	○	○	○
5	I had no problems navigating the game	○	○	○	○	○	○

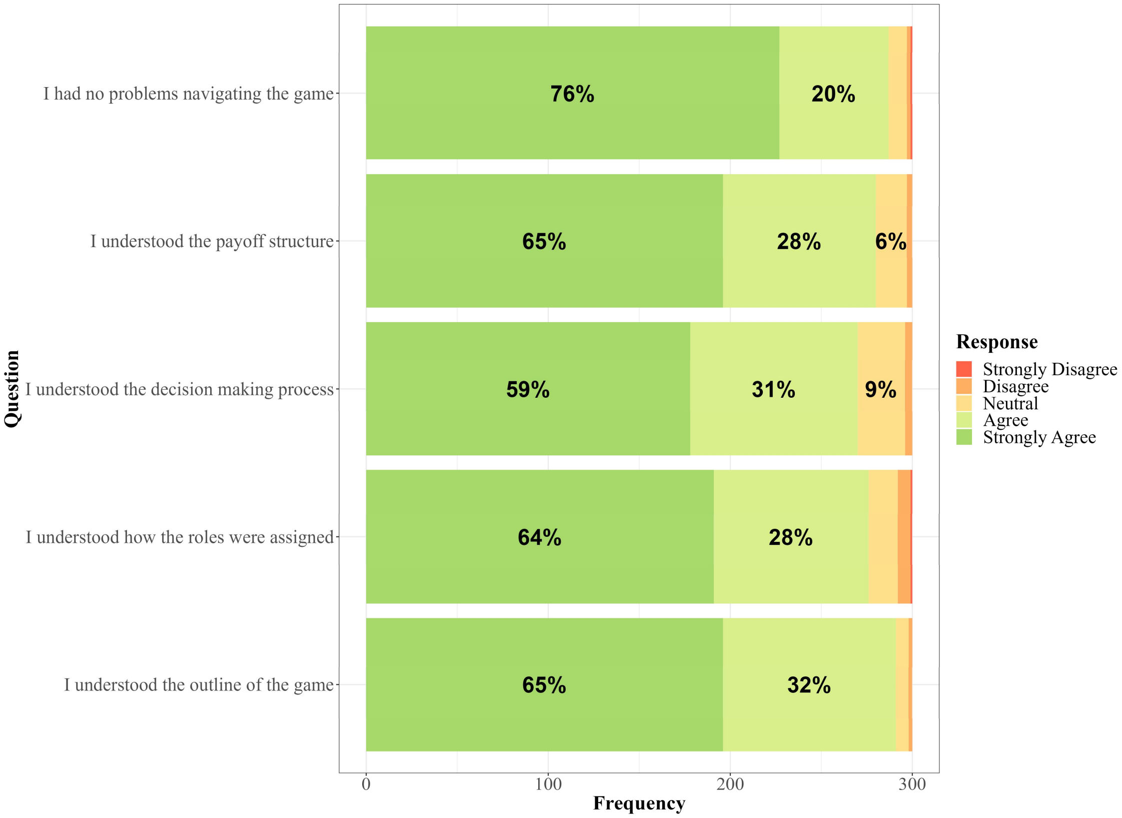
Appendix 6 – Post-Experiment AA Attitudes Survey

Table 11: Post-Experiment AA Attitudes Survey

	Please indicate the degree to which you agree/disagree with the following statements	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
1	I believe the automated agents to be a more fair inspector compared to a human	○	○	○	○	○	○
2	I believe the automated agents to make more objective decisions as an inspector	○	○	○	○	○	○
3	I preferred playing against a automated agents compared to a human	○	○	○	○	○	○
4	I would support the implementation of automated agents to control taxes by the government	○	○	○	○	○	○
5	I would support the implementation of automated agents to control wider areas of my life	○	○	○	○	○	○

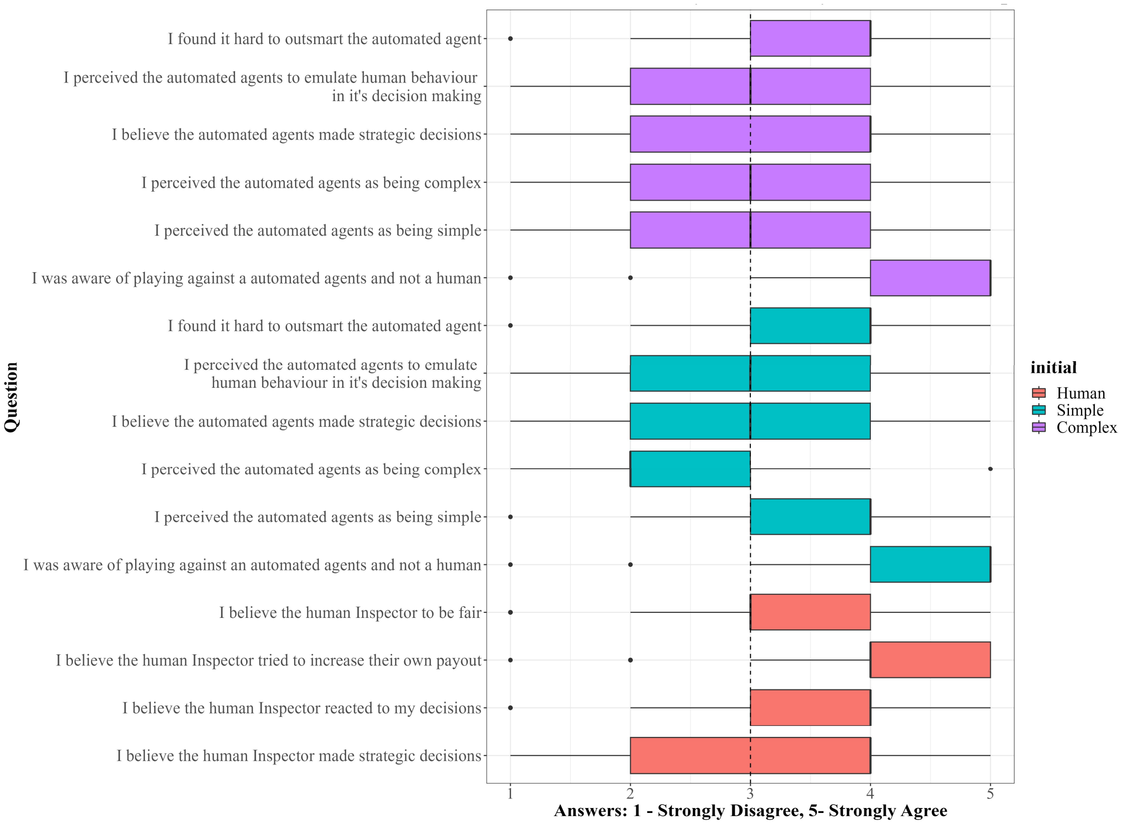
Appendix 7 – Post-Experiment Game Understanding Survey Answers

Figure 3: Post-Experiment Understanding of Experiment



Appendix 8 – Post-Decision Treatment Survey

Figure 4: Post-Experiment Treatment Survey



Appendix 9 – Boxplots ATI & Tax Measures

Figure 5: Boxplot showing the distribution of mean survey scores for the measures ATI and Tax Attitudes.

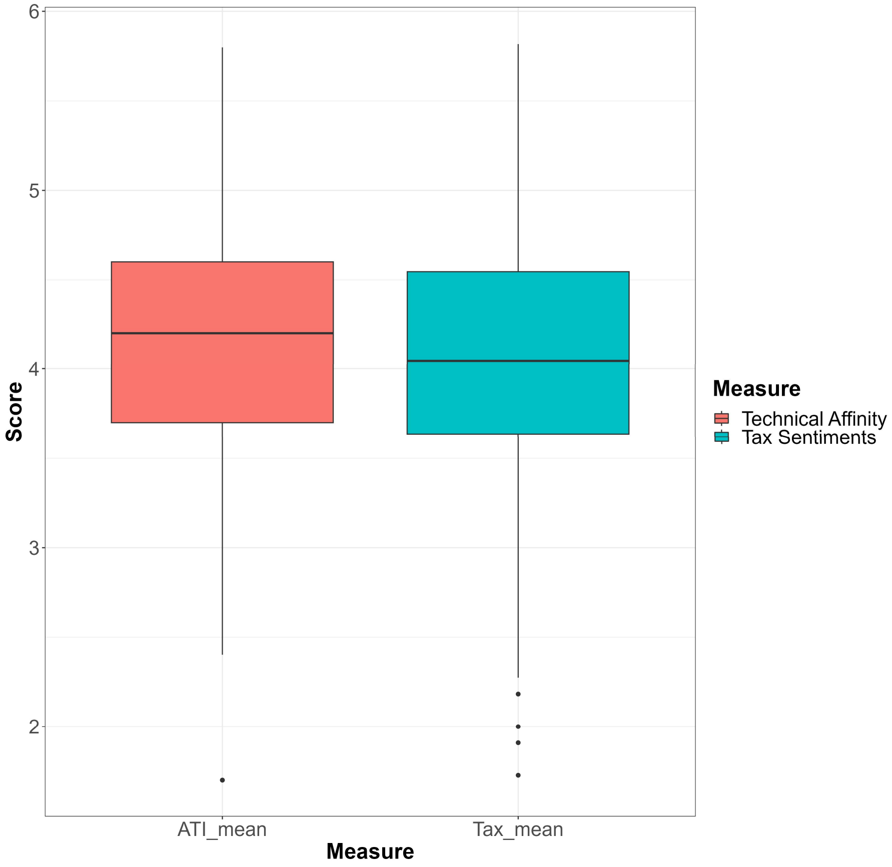
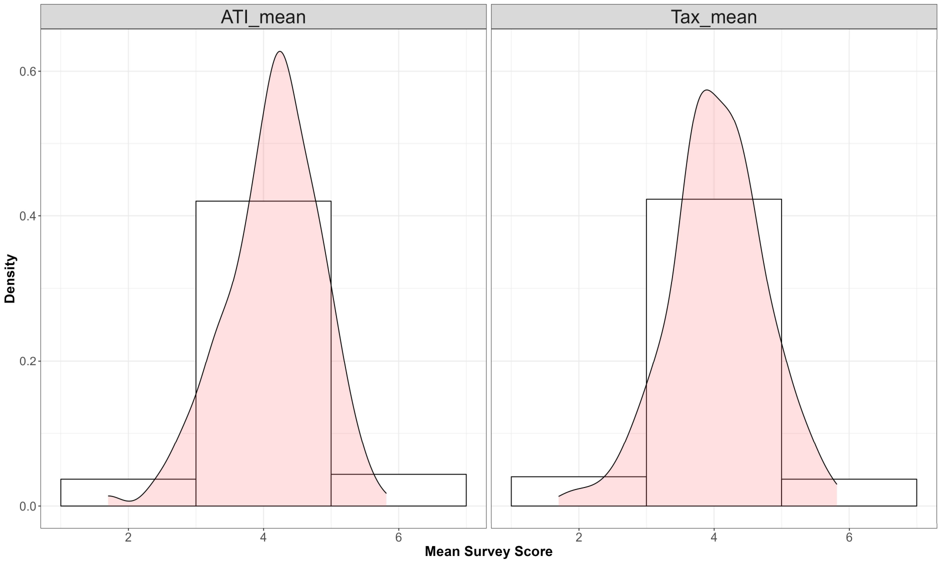


Figure 6: Histogram showing the distribution of mean survey scores for the measures ATI and Tax Attitudes.



Appendix 10 – Mixed effects logistic Regression Control Analysis

Table 12: Mixed effects logistic Regression with Control Variables

	Dependent variable: Evading Taxes						
	Sex	Age	Student Status	Employ-ment	Language	ATI	TAX
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Human Treatment (Ref: Complex Treatment)	-0.232***	-0.237***	-0.245***	-0.249***	-0.237***	-0.237***	-0.237***
	(0.048)	(0.048)	(0.052)	(0.053)	(0.048)	(0.048)	(0.048)
Simple Treatment (Ref: Complex Treatment)	0.167***	0.161***	0.150***	0.127**	0.161***	0.161***	0.161***
	(0.047)	(0.047)	(0.050)	(0.051)	(0.047)	(0.047)	(0.047)
Decision Sequence 0.4 (Ref: 0.53 mean)	0.458***	0.468***	0.541***	0.541***			
	(0.122)	(0.122)	(0.134)	(0.140)			
Decision Sequence 0.6 (Ref: 0.53 mean)	0.246**	0.249**	0.310**	0.323**			
	(0.120)	(0.119)	(0.132)	(0.134)			
Round	-0.012***	-0.012***	-0.014***	-0.013***	-0.012***	-0.012***	-0.012***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Caught in Previous Round	0.418***	0.416***	0.439***	0.443***	0.410***	0.410***	0.410***
	(0.049)	(0.049)	(0.053)	(0.054)	(0.049)	(0.049)	(0.049)
Male (Ref: Female)	-.0.068						
	(0.099)						
Age: 30-59 (Ref: <30)		-0.135					
		(0.112)					
Age: 60+ (Ref: <30)		-0.481					
		(0.855)					
Student (Ref: Non-Student)			0.161				
			(0.111)				
Employment: In Paid Work (Ref: Not in Paid Work)				-0.001			
				(0.121)			
Non-English primary Language (Ref: English)					-0.230**		
					(0.111)		
Technological Affinity Score						-0.022	
						(0.071)	
Tax Attitudes Score							-0.042
							(0.070)
Constant	-0.880***	-0.878***	-1.082***	-0.994***	-0.755***	-0.819***	-0.744**

	(0.113)	(0.103)	(0.124)	(0.117)	(0.124)	(0.309)	(0.296)
Observations	13,455	13,500	11,745	11,250	13,500	13,500	13,500
Log Likelihood	-8,088.112	-8,115.339	-6,975.257	-6,670.432	-8,114.089	-8,116.158	-8,116.028
Akaike Inf. Crit.	16,196.220	16,252.680	13,970.510	13,360.860	16,248.180	16,252.320	16,252.060
Bayesian Inf. Crit.	16,271.300	16,335.290	14,044.230	13,434.150	16,323.280	16,327.420	16,327.160

*Note: Significant noted as *p<0.1; **p<0.05; ***p<0.01. Mixed effects logistic Regression Model. Treatment reference category is complex treatment for all models. Decision Sequence is depicting the three difference inspection sequences with different mean inspection rates used for the inspection algorithm.. Rounds are round numbers 1-15 for each decision made within the round groups. All models span all round segments and have random intercept for both Participant ID as well as Participant and treatment sequence combinations (6 in total) to reflect within subject design considerations. Difference in n for models due to NA data for certain participants within that category.*

Appendix 11 – Round Effect Analysis

Figure 7 uses the aggregated data of the three round segments and shows the decisions both in a line. In the simple treatment we see a downward trend, while the complex and human treatment both exhibit a much more level trend. In Figure 8, we see that participants in the complex and simple treatment start with higher numbers of evade decisions, where participants in the human treatment exhibit lower evade decisions to start with. Overall, there are no clear outliers when looking at rounds. Table 13 shows a mixed effect logistic regression looking at the effect of round number on individual treatments as well as the interaction effects of treatment and round number We see a statistically significant decrease in the log-odds of evasion in the simple treatment (β = 0.021, p = <0.01), with complex treatment showing a non-significant negative and human a non-significant positive effect. When examining the interaction between treatment condition and round number, it can be observed that the effect of round number on the likelihood of evasion differed by treatment group, although this interaction was only marginally significant. The positive coefficient for the interaction term between human treatment and round number (β = 0.021, p = 0.053) indicates that for participants in the 'human' treatment group, the likelihood of evasion decreases less with each additional round relative to the complex treatment group. Conversely, the negative coefficient for the interaction simple treatment and round number (β = -0.019, p = 0.070) suggests a more pronounced decrease in evasion likelihood per round in the 'simple' treatment group compared to the complex treatment.

Figure 7: Mean of Evade Decisions by Round Nr. and Treatment (aggregated through all round segments)

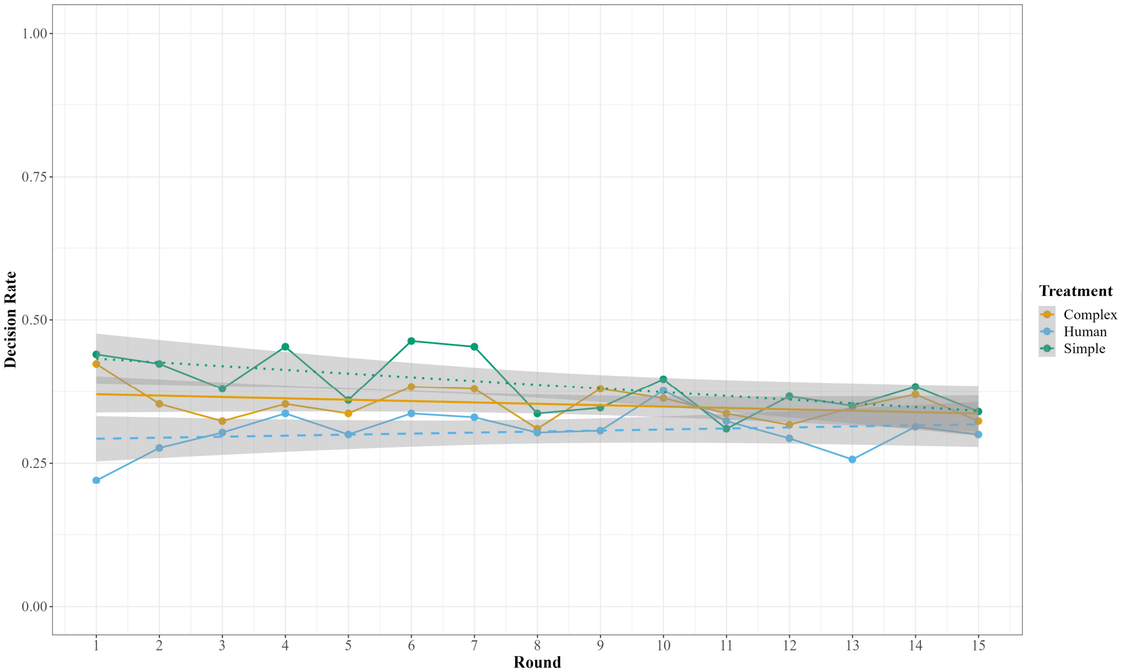


Figure 8: Decision Counts by Round number, grouped by Treatment (aggregated from all round segments)

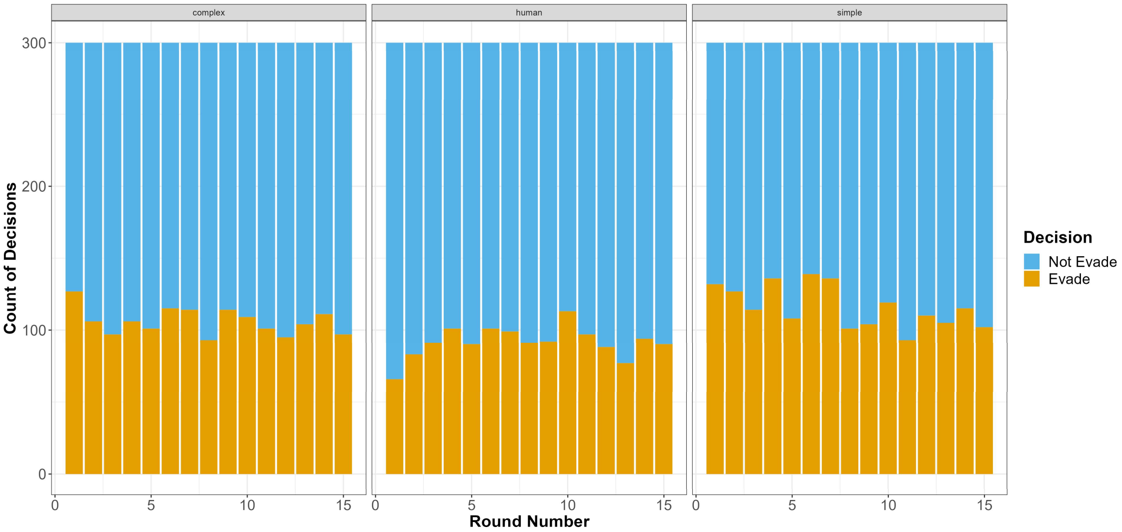


Table 13: Mixed Effects Logistic Regression for individual round effects and Round:Treatment Interaction Effects

	Dependent variable: Evading Taxes			
	Simple Treatment	Complex Treatment	Human Treatment	Interaction
	(1)	(2)	(3)	(4)
Round Nr.	-0.032*** (0.008)	-0.012 (0.008)	0.010 (0.008)	-0.012 (0.008)
Human Treatment (Ref: Complex Treatment)				-0.418*** (0.100)
Simple Treatment (Ref: Complex Treatment)				0.318*** (0.097)
Interaction Human Treatment:Round Nr. (Ref: Complex)				0.021* (0.011)
Interaction Simple Treatment:Round Nr. (Ref: Complex)				-0.020* (0.011)
Constant	-0.300** (0.140)	-0.616*** (0.164)	-1.070*** (0.101)	-0.609*** (0.088)
Observations	4,500	4,500	4,500	13,500
Log Likelihood	-2,838.011	-2,747.399	-2,613.833	-8,150.479
Akaike Inf. Crit.	5,684.021	5,502.798	5,235.665	16,316.960
Bayesian Inf. Crit.	5,709.668	5,528.445	5,261.313	16,377.040

Note: Significant noted as *p<0.1; **p<0.05; ***p<0.01. Mixed effects logistic Regression Model with binary evasion decisions for each round as dependent variable. Model 1-3 show the changes in log-odds for the decision to evade for each increase in round number grouped by treatment. Model 4 shows the interaction between treatment and round number. All models use complex treatment as reference group.

Appendix 12 – Initial Post Decision Treatment Analysis

Figure 9: Stacked Bar Plot for Post-Decision Survey Answers

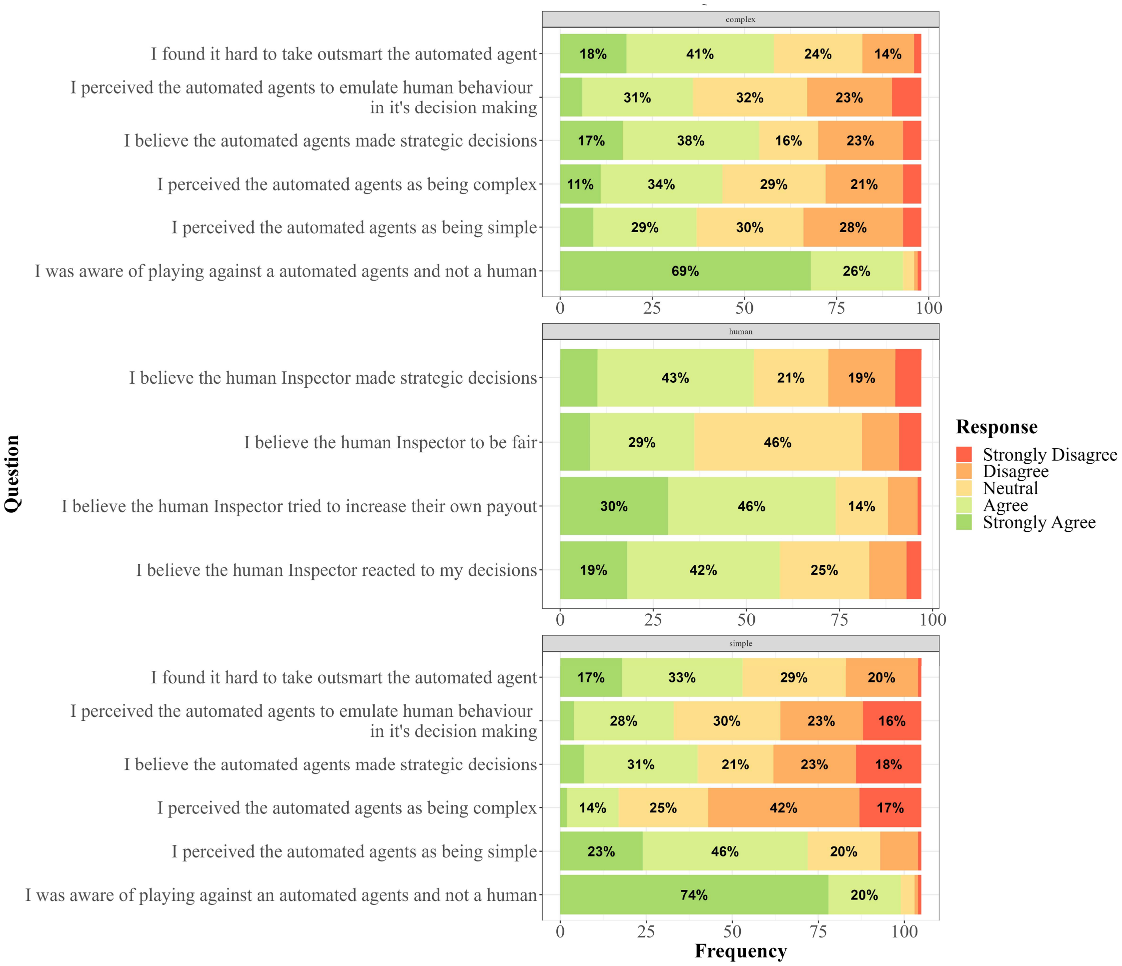


Table 14: Mixed Effects Logistic Regression Post-Decision Treatment Survey Effects

	Dependent variable: Mean Evasion			
	Simple Treatment	Complex Treatment 1	Complex Treatment 2	Human Treatment
	(1)	(2)	(3)	(4)
Mean survey score	0.149	-0.364	0.192	-0.295
	(0.590)	(0.434)	(0.353)	(0.416)
Awareness of AA	-0.203			
	(0.181)			
Simplicity of AA	-0.203			
	(0.155)			
Complexity of AA	-0.093			
	(0.165)			
Strategy of AA	-0.027			
	(0.145)			
AA Human Emulation	0.088			
	(0.142)			
Awareness of AA		0.150		
		(0.165)		
Simplicity of AA		0.071		
		(0.136)		
Complexity of AA		-0.053		
		(0.153)		
Strategy of AA		0.161		
		(0.141)		
AA Human Emulation			-0.095	
			(0.134)	
Difficulty Outsmarting AA			-0.103	
			(0.122)	
Human Strategy				0.138
				(0.162)
Reaction to Decisions				0.143
				(0.154)
Increasing own payout				0.228
				(0.139)
Constant	1.001	-0.214	-0.219	-1.572***
	(0.946)	(0.934)	(0.809)	(0.543)
Observations	1,575	1,470	1,470	1,455
Log Likelihood	-1,030.814	-974.110	-975.056	-907.859
Akaike Inf. Crit.	2,077.628	1,962.220	1,960.111	1,827.717
Bayesian Inf. Crit.	2,120.524	1,999.271	1,986.576	1,859.414

*Note: Significant noted as *p<0.1; **p<0.05; ***p<0.01. Mixed effects logistic Regression Model with Evasion Decision as dependent Variable. Models are split up in treatments, with Model 2 & 3 both denoting complex treatment, since model would not converge with all independent variables for complex treatment in one model.*

Full Questions for Table 14

Awareness of AA: I was aware of playing against an automated agents and not a human

Simplicity of AA: I perceived the automated agents as being simple

Complexity of AA: I perceived the automated agents as being complex

Strategy of AA: believe the automated agents made strategic decisions

AA Human Emulation: I perceived the automated agents to emulate human behavior in it's decision making"

Difficulty Outsmarting AA: I found it hard to take outsmart the automated agent

Human Strategy: I believe the human Inspector made strategic decisions

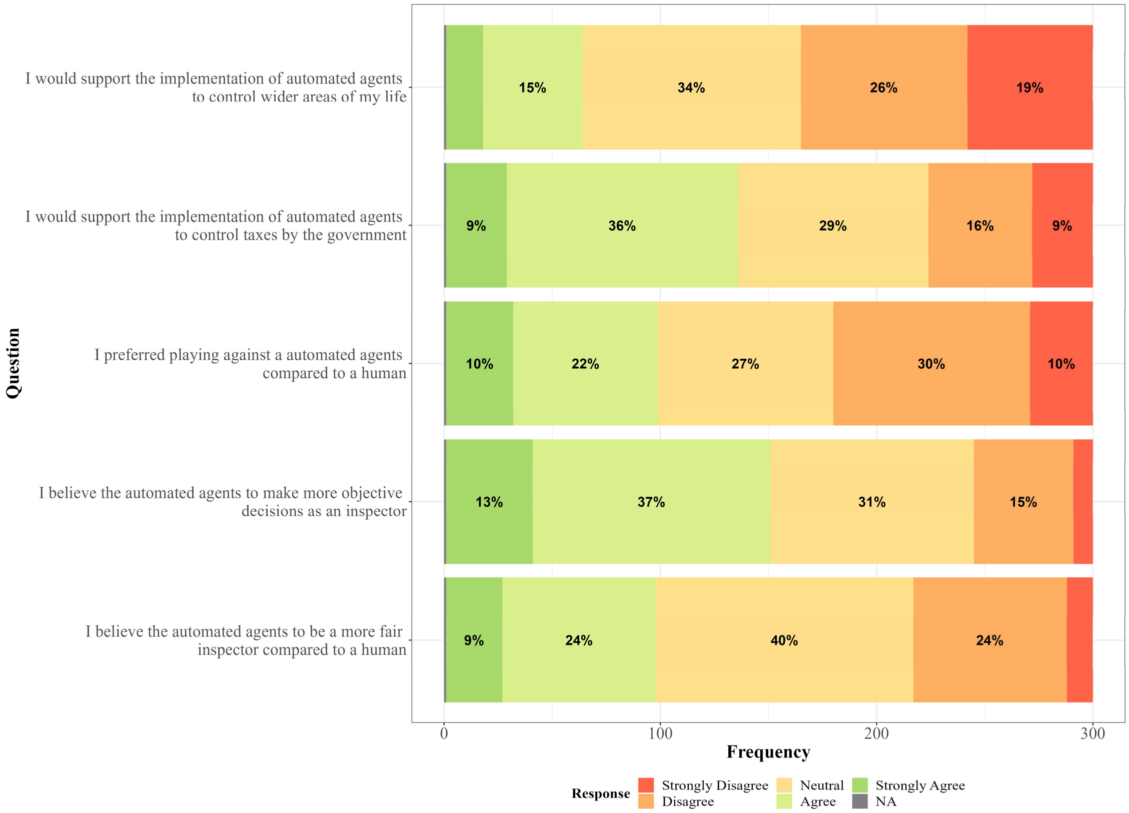
Reaction to Decisions: I believe the human Inspector reacted to my decisions

Increasing Own Payout: I believe the human Inspector tried to increase their own payout

Appendix 13 – AA Attitudes Analysis

All participants in the bot treatment filled out a post-experiment quiz which asked specific questions about the experience of playing against a bot and overall attitudes towards AAs. The questions were asked with a 1-5 Likert Scale, with 1 indicating strong agreement and 5 indicating heavy disagreement. Figure 10 shows the answers of 6 questions concerning overall attitudes towards AAs, with most answers showing quite a balanced picture. There's a higher concentration of disagreement concerning the implementation of AAs to control wider areas of one's life (19% Strongly Disagreee, 26% Disagree), but people would be more willing to support AAs to control tax reports (9% Strongly Agree, 36% Agree). People slightly favored playing against a human compared an AA, with 10 % Strongly disagreeing and 30% disagreeing to having preferred playing against an AA compared to human. On the other hand, participants thought AAs make more objective decisions as inspectors (13% strongly agree, 37% agree). Yet. They only slightly deem AAs to fairer inspectors compared to human, with 5% higher strong agreement compared to strong disagreement.

Figure 10: Stacked Bar Chart for Sentiments towards Automated Agents



In a next step, these questions were analyses in their effects on the mean decision to evade through a linear regression model. A linear regression model was chosen here as it allows for the inclusion of more variables without encountering convergence issues. Both the individual answers as well as the mean score across all questions were used as independent variables. Table 15 shows the results of the linear regression. Model 1 shows the overall effect of the survey questions on the mean evasion rate, while model 2-4 show the effects for the individual treatments. Initial treatment independent variables were used to account for the within-subject design effects. We can see that the mean score of the survey does not significantly affect mean evade decisions. Concerning individual questions, we see no significant effect except for the question “I would support the implementation of AAs to control wider areas of my life”, where we see a moderately significant effect (coefficient 0.036, p-value < 0.05). This means the higher the disagreement with this statement, the higher the mean evasion amount of the participant. This effect holds for the complex treatment (coefficient 0.037, p-value < 0.05, and becomes even more significant in the human treatment (coefficient 0.049, p-value < 0.01). The influence of the initial treatment is also significant in the complex and simple treatment decisions. Overall, preference of opponent, fairness and objectiveness sentiments did not seem to have an influence, but the acceptance of AA to control overall life aspects seemed to influence people’s strategic decisions.

Table 15: Linear Regression with AA Sentiments

	Dependent variable: Mean Evasion			
	Overall	Complex Treatment	Simple Treatment	Human Treatment
	(1)	(2)	(3)	(4)
Mean Post-Survey Score	-0.028	-0.053	-0.023	-0.029
	(0.045)	(0.053)	(0.055)	(0.052)
Initial Treatment Human		-0.013	0.120***	-0.071**
		(0.029)	(0.030)	(0.028)
Initial Treatment Simple		0.149***	0.071**	-0.042
		(0.029)	(0.030)	(0.029)
Fairness	0.009	0.021	-0.012	0.022
	(0.016)	(0.019)	(0.020)	(0.019)
Objectiveness	-0.014	-0.021	-0.003	-0.018
	(0.016)	(0.019)	(0.019)	(0.018)
Support AA Tax Inspection	0.002	0.015	0.012	-0.013
	(0.015)	(0.017)	(0.018)	(0.017)
Support AA broadly	0.036**	0.037**	0.022	0.049***
	(0.015)	(0.018)	(0.018)	(0.017)
Constant	0.355***	0.331***	0.342***	0.337***
	(0.043)	(0.054)	(0.056)	(0.053)
Observations	300	300	300	300
R^2	0.038	0.141	0.069	0.068
Adjusted R^2	0.022	0.121	0.046	0.046
Residual Std. Error	0.172 (df = 294)	0.203 (df = 292)	0.212 (df = 292)	0.201 (df = 292)
F Statistic	2.319** (df = 5; 294)	6.872*** (df = 7; 292)	3.071*** (df = 7; 292)	3.063*** (df = 7; 292)
Note:			*p<0.1; **p<0.05; ***p<0.01	

Universität Zürich UZH

Department of Sociology

Declaration of Authorship for Academic Work

Title of the Work*: **Evading the Algorithm: The Increased Propensity for Tax Evasion in Human-Computer Interactions**

Module Name: **Master Thesis**

Supervisor/Lecturer: **Dr. Prof. Heiko Rauhut**

* only main title necessary – no subtitles

I hereby expressly declare that the paper submitted by me is an original work written by myself in my own words without any unauthorized assistance. If the paper is by more than one author, I hereby confirm that the parts of the paper written by authors other than myself are correctly and clearly identified as such and are explicitly attributed to the respective author(s).

I additionally confirm that this paper – as a whole or in parts – has not previously been submitted to the University of Zurich or to any other university or educational institution for the purpose of obtaining a degree or academic credit, nor will it be submitted again in the future for the purpose of obtaining a degree or academic credit.

Use of Sources

I expressly declare that all references to external sources (including tables, graphics, etc.) contained in the abovementioned paper have been identified as such. In particular, I confirm that without exception and to the best of my knowledge, I have properly indicated the authorship of borrowed verbatim statements (quotations) and statements by other authors reproduced in my own words (paraphrases).

Sanctions

I acknowledge that work which violates the principles of the declaration of authorship – particularly work containing quotations or paraphrases without indication of origin/source – is considered plagiarism and may result in corresponding legal and disciplinary consequences (pursuant to §§ 10ff of the Disciplinary Regulations of the University of Zurich, order 415.33, and in accordance with § 12 of the Framework Ordinance for Studies in the Bachelor's and Master's Programs of the Faculty of Arts and Social Sciences of the University of Zurich, order 415.455.1).

With my signature, I confirm the accuracy and veracity of the information provided in this declaration.

Author: Nico Mutzner

Student ID #: 21-709-811

Location and Date: ZH, 15.05.2023

Signature: N.