TYPE Original Research
PUBLISHED 05 October 2023
DOI 10.3389/frbhe.2023.1227166

OPEN ACCESS

EDITED BY
Nicola Lacetera,
University of Toronto, Canada

REVIEWED BY
Ksenia Keplinger,
Max Planck Institute for Intelligent Systems, Germany
Enrico Petracca,
University of Bologna, Italy

*CORRESPONDENCE
Nico Mutzner
✉ nico.mutzner@uzh.ch

RECEIVED 22 May 2023
ACCEPTED 14 September 2023
PUBLISHED 05 October 2023

CITATION
Mutzner N, Oberhauser V, Winter F and Rauhut H (2023) Evading the algorithm: increased propensity for tax evasion and norm violations in human-computer interactions. *Front. Behav. Econ.* 2:1227166. doi: 10.3389/frbhe.2023.1227166

Evading the algorithm: increased propensity for tax evasion and norm violations in human-computer interactions

Nico Mutzner*, Vincent Oberhauser, Fabian Winter and Heiko Rauhut

Department of Sociology, University of Zurich, Zurich, Switzerland

Today's modern world is characterized by an increasing shift from human-to-human interaction toward 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 our understanding of HCI is evolving, it is still in nascent stages. This is particularly true in the sphere of non-cooperative strategic interactions between humans and automated agents, which remains largely unexplored and calls for further investigation. 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. Here, 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 they are inspected by automated agents rather than by humans. These results may also be transferred to norm violations more generally, which may become more likely when participants are controlled by computers rather than by humans. Our results further show that participants were less likely to evade taxes when playing against an automated agent described as a complex AI, compared to an automated agent described as a simple algorithm, once they had experienced different agents.

KEYWORDS

inspection game, human-computer interaction, tax evasion, decision-making, humans vs. machines, automated agent, crime, social norms

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). Automated Agents (AAs), although typically embedded in software, are considered physical in the broader sense, as they exist within and interact with our physical world, often through digital devices or machinery, and thus affect tangible outcomes in a defined environment. 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 toward 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 (Lohr, 2018; Leslie, 2020). Very recent discussions about the new natural-language models GPT-3 and GPT-4, and their application in ChatGPT (OpenAI, 2022, 2023), 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 affect 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 experiments 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 prisoner's dilemma, dictator games, ultimatum games, negotiation games, and public-goods games (Kiesler et al., 1996; De Melo et al., 2016; Weiss et al., 2020; Lee et al., 2021; Nielsen et al., 2022), 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 as representing social-interaction effects well (Falk and Fischbacher, 2002) and lending themselves well to being 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, which constitutes a novel approach to this type of experiment. 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? Our design allows us to identify how the deployment of different agents impacts participants' strategic decision-making depending on whether they perceive their counterpart as being a human, a simple AA, or a complex AA. 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. With this in mind, our study expands upon current research by incorporating AAs in the inspection game, particularly within the context of a repeated mixed strategy approach. Previous experiments predominantly utilized one-shot games when implementing AAs, providing a limited view of strategic interactions. Our research, however, reveals a more intricate facet of these interactions in non-cooperative games. By introducing sequentiality and mixed strategies, we provide a richer understanding of the dynamics underlying deviant strategic decision-making in human-computer interaction. This approach not only fills an existing gap in the literature, but also adds a new dimension to the discourse on strategic decision-making within human-computer interaction.

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 simple 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 the later parts of the experiment. We further find significant round effects, where participants' evasion probabilities would either reduce or increase over the 15 rounds played, depending on the agent type. 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 fewer norm deviations 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 toward 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. For instance, given the increasing digitization of taxation and financial systems worldwide, governments and tax authorities stand to benefit greatly from this research. An understanding of how people respond differently to automated agents could inform the design and deployment of digital tax platforms. Likewise, developers and companies producing the automated agents, such as AI-based inspection or auditing software, would gain insight into how the perceived complexity of their technology influences user behavior and compliance. Therefore, it is evident that the findings of this study are not only relevant to researchers in the field, but also to these stakeholders—policymakers, tax authorities, and technology developers—, who need to comprehend these interactions for the creation of effective and user-friendly systems.

2. Related literature

2.1. Experimental approaches to crime

Becker (1968) first introduced an economic approach toward 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 delinquency within a rational choice framework. The original 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. These versions employ a more strategic decision-based approach to crime, expanding on the rational-choice framework. Further, 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 these limitations also shifts the focus away from the previous approaches, which rely heavily on a more economic model of the inspection-game concept, instead moving it more toward 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 et al. (1990) have pointed out, the one-shot nature of the model 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 social norms, where inspectees 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. For the development of the experimental design of our study, the first inspection-game experiments by Rauhut (2009, 2015) inspired our design choices.

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 the 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 decision foundation found in inspection games (Mascagni, 2018). Determining factors of tax compliance within these experiments are based both on economic models (Beck et al., 1991) as well as on 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 and 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 cost in terms of feeling "bad" for cheating, which can be placed in a social context, 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 and 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 in the inspection literature have focused mostly on the taxpayer or inspectees 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 that 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 et al. (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 various types of social behavior 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 to 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 and Kuchenbrandt, 2012). Computers can also be seen as teammates, where humans who 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; Salem et al., 2015; Robinette et al., 2016).

Nass and Moon (2000) explain the existence of these social attributions onto computers on the basis of Langer's (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 confirm 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. Further, as Chugunova and Sele (2022) note in their literature review of HCI experiments, certain ascriptions of social characteristics such as emotional responses can be less prominent when interacting with computers, as well as lower extents of social attribution being dependent on characteristics of the automated agent. Nevertheless, they conclude that social concerns and notions do take place, and research into these areas of interaction is therefore necessary to investigate further how such notions can affect the decisions humans make when placed with or against a computer in specific situations.

Historically, the inclusion of AAs to test such considerations has been scarce. For the inspection game, AAs have mostly been used as automated tools to simulate decisions of rational learning models (Rauhut and Junker, 2009; Rauhut and Jud, 2014), or multi-agent systems used for automated negotiations (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 to include them as active players in the interaction scenarios. One of the first examples of including computers in an experiment was Kiesler et al. (1996), who confirmed the 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 so less 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 making 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 (Schniter et al., 2020; Weiss et al., 2020; Nielsen et al., 2022). 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). In addition, people were more dishonest toward AAs, compared to humans in a coin-toss task (Maréchal et al., 2020). Therefore, while people go against theoretical expectations of rationality and selfishness and treat AA's socially, they still tend to be less prosocial and less honest with computers and exploit them more compared to interactions with 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 et al. (2021) looked at how the modeling 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 that 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 (Daylamani-Zad and 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 Ii-intelligence.*

Our distinction between a simple algorithm and a complex AI may implicitly invoke concepts of explainability in AI, pertaining to the transparency of an AI's decision-making process (Xu et al., 2019). However, we chose the more straightforward terms of simple vs. complex for the sake of clarity and simplicity and its direct interpretation of complexity. While we do not explicitly engage with the discourse on explainability, our experimental design does incorporate elements of decision-making explanations, so that we indirectly contribute to this strand of the literature.

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 the inspector. The game consists of three segments of rounds, each with 15 decisions, and an initial endowment of 100 tokens for each round segment. In each round, participants can decide either to underreport their taxes or to 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 ten tokens. If the participant decides to underreport, but does not get audited, they receive a payment of five tokens. To ensure symmetry in decisions, the same payoffs are used for the inspector, meaning a successful inspection results in a five-token reward, while an inspection on a full report leads to an inspection cost of ten tokens. If both players are in a situation of no underreporting and no inspection, no balance change occurs. The formula for the game-theoretical probability for evading taxes is $S_i^* = \frac{k}{r}$, with S_i^* denoting the probability of committing tax evasion, k denoting the inspection cost, and r denoting the inspection reward (Rauhut, 2015). With the parameters used in this experiment, this results in a mixed-strategy equilibrium, with a 0.5 probability for tax evasion. Due to these parameters, there is no equilibrium in pure strategies, and participants are forced to determine their action for each round strategically. 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. Each decision sequence was randomly assigned to participants in each treatment, with the aim of achieving an equal distribution across them. This means each decision sequence was equally likely to appear in any given treatment. The decision sequences are used as control variables to ensure that 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 behavior under controlled conditions.

TABLE 1 Payoff structure for taxpayer (participant).

		Inspector	
		Audit	No audit
Taxpayer	Underreport	$-10, 5$	$5, 0$
	Fully Report	$0, -10$	$0, 0$

3.2. Treatment

This study uses a within-subject design, featuring three treatments. The initial 15 rounds can be treated as a between-subject design, providing an overview of the initial exposure to each treatment. The subsequent two-round segments follow a within-subject design, enabling comparability and facilitating causal inference. Each treatment is 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 will play against a complex AI that mimics human decisions. Importantly, the inspectors actually do not differ, and the inspector plays out the same pre-defined decision sequences for all treatments. An alternative approach could be that participants are informed that they have a certain probability of encountering a specific agent. With such an approach, we would avoid participant deception. However, due to the online nature of our sample, we have decided to forego such a procedure and use this straightforward and simple approach for perception manipulation. This ensures that treatments only differ by the distinct perception of the inspector, and not by underlying differences in decision-making by humans and AA. 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 separate 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 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 were aware that their counterpart was either human or a computer, in order 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 on top of each decision page whom they are playing against. Upon the conclusion of the experiment, participants are informed about the manipulation of perception that took place in the experiment through a debriefing 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 toward AI and machines have shown that sociodemographic factors, technical affinity, as well as knowledge of technology are critical factors that can influence perceptions on AA implementations. Examples include applications of AI in healthcare (Fritsch et al., 2022), AI in decision-making (Kushwaha et al., 2022), general attitudes toward AI (Zhang and Dafoe, 2019; Selwyn et al., 2020), and different forms of automated system applications (Langer, 1992). Therefore, this experiment elicits such factors through several survey questionnaires. First, participants complete a nine-item questionnaire concerning their technical affinity. The nine items are based upon the Affinity for Technology Interaction (ATI) scale by Franke et al. (2019), which is measured on a six-point Likert scale from "Completely disagree" to "Completely agree" (Supplementary Table 1). Second, they fill out a survey about their attitudes toward taxes which was built on segments from the Comprehensive Taxpayer Attitude Survey (2021) (see Supplementary Table 2). Both measures are treated as additive indices ranging from 1 to 6. Attitudes toward taxation have been shown to be an influencing factor on decision-making in tax evasion, and therefore warrant inclusion as a control variable (Wärneryd and Walerud, 1982; Torgler, 2002). To ensure understanding of the game, treatment effectiveness, and gauge overall attitudes toward AAs, participants complete three additional surveys. After the first 15 decisions, participants are asked about their experience of playing against their specific treatment (Supplementary Tables 3, 4). 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 are clear (Supplementary Table 5). Lastly, participants are asked about their attitudes toward AAs in general, as well as the use of AAs to control taxes and wider aspects of their life (Supplementary Table 6). 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 and experiment

The platform used to program the experiment itself was O-Tree (Chen et al., 2016), using Python. 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. 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 (Peer et al., 2017; Palan and Schitter, 2018). 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 and 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. No problems were encountered while the experiment was conducted, and the data were collected on a weekday afternoon, when high participation rates are usually observed. The required participant count of 300 was calculated through a power analysis of pilot data. The payment for participants was based on a fixed fee (2£ for 20 min), plus a variable bonus based on performance. The median time for completion of the experiment was 19 min and 18 s, 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 would be calculated from one of the three round segments, which would be randomly chosen as the payoff-relevant round segment at the end of the experiment. Participants were informed that their data would be kept anonymously, since the data were anonymized for the analysis.

4. Results

4.1. Descriptive

The experiment reached a final participant count of 300 individuals. Overall, we achieved a high understanding of the experiment, which was tested with the post-experiment survey asking participants about their understanding of different components of the experiment (Supplementary Figure 1). 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 through the game. Concerning sociodemographic attributes, individual information about participants was acquired through the user data from Prolific.

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 countries	South Africa	68	22.67%
	Poland	58	19.33%
	Portugal	57	19.00%
	Italy	20	6.67%
	Greece	14	4.67%

Individual variables were then grouped into larger groups to facilitate analysis, with NA designated to data that had expired. Table 2 gives an overview of the used categories. Regarding sex, a slight skew toward male participants was recorded, with 53.7% (n = 161) of the participants identifying as male and 46% (n = 138) as female, with one person not wishing do disclose their sex. For other characteristics, the sample indicates that a majority of participants were under 30 (73%, n = 219), in paid work (57%, n = 171), predominantly white (67%, n = 201) and having 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 with 19.33% (n = 58), and Portugal with 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. Thirty-seven countries were represented in the sample, although many were only comprised of a single participant. An examination of the distribution of the Tax and ATI measures reveals that both variables predominantly adhere to a normal distribution (Supplementary Figure 2). This suggests that they are unlikely to introduce issues in the statistical analyses due to skewness or outliers.

4.2. Decision analysis

The post-decision survey after the first 15 rounds was used to check if treatment manipulation was successful. Participants overall recognized playing against an AA compared to a human in both AA treatments, with 94% agreeing that they were playing against an AA in the simple treatment, and 95% in the complex treatment. The AA treatments were also perceived differently, with 31% more agreement toward estimating the AA to be simple in the simple treatment compared to the complex treatment, and 29% more agreement that the AA was perceived as complex in the complex treatment compared to the simple treatment. This suggests that our treatment manipulation was successful, with participants perceiving the same decision algorithm as being different between the treatments (see Supplementary Figure 3).

Table 3 and Figure 1 present the mean evasion rates, count, and standard deviation across treatments and round groups. In the first 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 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.

A mixed-effects logistic regression model was estimated to leverage the binary decision design used within this experiment; results are shown in Table 4. This method aptly handles binary-decision data, such as the decision to evade or not to evade taxes, while also accommodating for the nested nature of our observations considering decisions grouped by 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 = −0.536, $p < 0.01$) and across all rounds (coefficient = −0.242, $p < 0.01$), 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 31–45 (coefficient = 0.359, $p < 0.1$). This might be reflective of the decrease of the evasion rates in the complex treatment found in the later round segments, while the evasion rates in the simple treatment stayed more consistent. Over all rounds, the simple treatment showed a significant positive effect on evasion (coefficient = −0.158, $p < 0.01$). This supports 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 in addition to the treatment variables were incorporated. The decision sequence, denoting the inspection rate of the automated inspector, showed a significant effect across rounds

TABLE 3 Table of evasion decisions and descriptive statistics.

	Complex			
	All rounds	Rounds 1–15	Rounds 16–30	Rounds 31–45
Mean evasion	0.35	0.46	0.33	0.28
SE	0.05	0.09	0.10	0.09
CI low	0.33	0.42	0.28	0.25
CI high	0.38	0.50	0.37	0.32
Count	300	98	99	103
	Human			
	All rounds	Rounds 1–15	Rounds 16–30	Rounds 31–45
Mean evasion	0.31	0.34	0.30	0.28
SE	0.06	0.08	0.10	0.11
CI low	0.28	0.31	0.26	0.24
CI high	0.33	0.38	0.34	0.32
Count	300	97	102	101
	Simple			
	All rounds	Round 1–15	Rounds 16–30	Rounds 31–45
Mean evasion	0.39	0.44	0.37	0.34
SE	0.05	0.09	0.08	0.11
CI low	0.36	0.40	0.34	0.30
CI high	0.41	0.48	0.41	0.39
Count	300	105	99	96

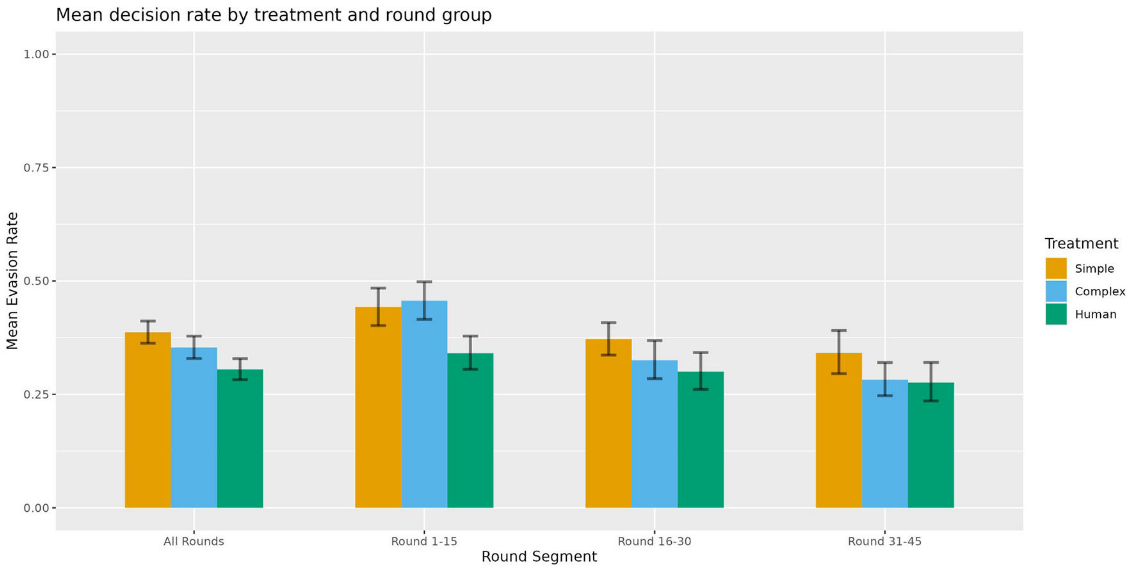

FIGURE 1
Mean decision rate by treatment and round group.

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 31–45 (coefficient = 0.482, $p < 0.05$), in comparison to the reference category of 0.5 mean inspection rate. Yet, the 0.43 decision sequence was not significant in the other round segments, nor over all rounds (coefficient 0.182, $p > 0.1$). When considering the decision sequence with a 0.63 mean inspection rate, we also see

TABLE 4 Mixed-effects logistic regression model of tax evasion decisions.

	Mixed effects logistic regression model all rounds			
	Dependent variable: evade			
	All rounds	R: 1–15	R: 16–30	R: 31–45
	(1)	(2)	(3)	(4)
Human treatment (ref: complex treatment)	−0.242*** (−5.01)	−0.536*** (−3.82)	−0.167 (−1.12)	−0.0519 (−0.30)
Simple treatment (ref: complex treatment)	0.158*** (3.36)	−0.0484 (−0.35)	0.224 (1.53)	0.359* (2.10)
Decision sequence 0.6 (ref: 0.5)	−0.283* (−2.25)	0.0248 (0.18)	−0.481*** (−3.30)	−0.472** (−2.79)
Decision sequence 0.4 (ref: 0.5)	0.179 (1.40)	0.124 (0.90)	0.00101 (0.01)	0.482** (2.88)
Round	−0.0181*** (−12.02)	−0.00372 (−0.50)	−0.00540 (−0.69)	−0.0300*** (−3.61)
ATI mean score	−0.0433 (−0.56)	−0.00781 (−0.09)	−0.0246 (−0.28)	−0.125 (−1.21)
Tax attitude mean score	−0.0540 (−0.74)	−0.0217 (−0.27)	−0.0246 (−0.29)	−0.0433 (−0.45)
Constant	0.520 (0.96)	0.431 (0.74)	0.0221 (0.34)	0.765 (0.98)
Control	Yes	Yes	Yes	Yes
Observations	13,455	4,485	4,485	4,485
Subjects	299	299	299	299

Significant noted as $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Mixed-effects logistic regression model. Treatment reference category is the complex treatment for all models. Decision Sequence depicts the three different 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. Control indicates that analysis has been run with control variables for sex, education, language, student status, ethnicity, and employment group, with no significant changes in main effects. All models have a random intercept for both participant ID as well as participant and treatment sequence combinations (six in total) to reflect within-subject design considerations.

a positive significant effect in rounds 16–30 (coefficient 0.481, $p < 0.01$) and rounds 31–45 (coefficient 0.472, $p < 0.05$), as well as overall rounds (coefficient 0.283, $p < 0.1$). Therefore, the decision sequence showed a positive effect on the participants' decision to evade taxes. Furthermore, the round variable shows a significant negative effect in rounds 31–45 (coefficient $= -0.03$, $p < 0.01$) and across all rounds (coefficient $= -0.0181$, $p < 0.01$). This indicates that, with each round within the 15 round segments, the log-odds of evading decreases by -0.0181. The additional measures of mean survey scores for tax attitudes and technological affinity both did not show significant effects. For simplicity and clarity in the analysis and to facilitate a better understanding and interpretation of the results, other control variables that were analyzed are excluded in the final analysis. While the control variables had slight impacts, the significance of the main effects persisted throughout all variations of the model.

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 4, which represents changes within decision-making in each round, exhibits an overall significant effect. To understand how rounds affect participants' decisions in more detail, the individual round segments have been aggregated to examine overall trends of evasion decisions over the 15 rounds for each treatment (Supplementary Figure 4). 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 (Supplementary Table 8), where the negative effect of the 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 with a linear-probability model. We find a significant interaction effect for both simple and human treatment with round numbers compared to the complex treatment (See Supplementary Table 9). The positive coefficient for the interaction term between human treatment and round number (coefficient $= 0.04$, $p < 0.1$) 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 between simple treatment and round number (coefficient $= -0.04$, $p < 0.1$) 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 (Supplementary Figure 3). 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 the decision to evade taxes as the dependent variable (Supplementary Table 7). 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 toward AAs and the use of AAs as inspectors (Supplementary Figure 5). Using these variables, a linear regression was run using the mean decisions to evade taxes as a dependent variable as well as the mean survey score and the individual answers as independent variables (Supplementary Table 10). 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 for participants in the human treatment. These additional results confirm the robustness of our main findings.

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 confirmed by our data: 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 (De Melo et al., 2016; Maréchal et al., 2020), 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 (Salem et al., 2015; Robinette et al., 2016), and therefore see higher risk associated with evading taxes.

From an experimental perspective, it is important to remark on the internal variables that also influenced decisions. We saw that the decision sequences used by the automated bot inspectors had a significant influence on the decisions of participants. It can be presumed that decision sequences with higher inspection rates lead to more catches of evasions, influencing the participants' decisions to evade. 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 confirmed the robustness of our results. Most sociodemographic factors did not show any significant influence. 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 the ATI literature and in the tax literature, where such sentiments were deemed to have an influence on decisions and interactions (Wärneryd and Walerud, 1982; Torgler, 2002; Franke et al., 2019).

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 et al. (1994) and Nass and Moon (2000), suggesting that humans can perceive machines to emulate human behavior and strategy, which in turn influences their behavior toward 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 toward interacting with humans over machines in social decision-making scenarios, which has previously been identified in other studies (McCabe et al., 2001; Gallagher et al., 2002). 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 toward 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 AA's capabilities and subsequent risk estimations.

Our study does not come without limitations. Firstly, the study population was recruited from an online platform, which has a higher likelihood of consisting of participants who have higher technological expertise and positive viewpoints of technology compared to the general population, 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 and 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 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.

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. First, from an economic-strategic decision perspective, this paper confirms the fact that different agents elicit different risk and reward perceptions, altering the outcome of strategic decision-making previously found in experiments with humans (Rauhut, 2015). Second, the convergence of complex and human inspectors in later rounds might indicate that, if individuals perceive automated agents as capable of mimicking human control agents, they may apply the same social scripts and engage in similar strategic decisions as they would with human control agents. Third, we confirm that descriptions of complexity alter how AAs are perceived, and consequently how humans make decisions against them. Therefore, this paper outlines the importance of how the perception of a control agent shapes human strategic decision-making, and how the manipulation of such perceptions can significantly change decision outcomes. Our study's practical implications extend to various sectors. Policymakers could use our findings to balance the economic benefits of Automated Agents (AAs) against the risk of increased norm-deviating behavior. Developers might enhance AI system transparency about complexity to influence user behavior positively and foster compliance. Our results also highlight the importance of ethical guidelines for AA usage and portrayal, particularly when these can significantly sway individual and societal behavior. Furthermore, organizations integrating AAs could leverage our insights to develop proactive strategies, mitigating norm-deviation risks and promoting better operational efficiency and compliance. Future research might expand on these findings in other contexts, while also addressing personal preferences and strategic estimations of humans 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-experimental survey, while people might rate computers as being more objective and fair, 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.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Committee of the Faculty of Arts and Social Sciences, University of Zurich, number 23.04.22. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

NM contributed to the conception and design of the study, analyzed the data, and wrote the initial manuscript. VO conducted the online experiment and processed the data. All authors contributed to the experimental design, data analysis, manuscript revision, and approved the submitted version.

Funding

This work was supported by Grants BSSGI0_155981 and 10001A_176333/1 from the Swiss National Science Foundation.

Acknowledgments

We are grateful for the suggestions and insights from Alexander Ehlert, Justus Rathmann, Antonia Velicu, Debora Frei, Strahinja Popovic, Diego Strassman, and Hagen Worch.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frbhe.2023.1227166/full#supplementary-material

References

Alm, J. (2012). Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. *Int. Tax Public Finan.* 19, 54–77. doi: 10.1007/s10797-011-9171-2

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

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

Beck, P., Davis, J. S., and Jung, W. O. (1991). Experimental evidence on taxpayer reporting under uncertainty. *Account. Rev.* 66, 535–588.

Becker, G. S. (1968). Crime and punishment: an economic approach. *J. Polit. Econ.* 76, 169–217. doi: 10.1086/259394

Bianco, W. T., Ordeshook, P. C., and Tsebelis, G. (1990). Crime and punishment: are one-shot, two-person games enough? *Am. Polit. Sci. Rev.* 84, 569–586. doi: 10.2307/1963536

Blumenthal, M., Christian, C., and Slemrod, J. (2001). Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. *Natl. Tax J.* 54, 125–138. doi: 10.17310/ntj.2001.1.06

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree – An open-source platform for laboratory, online, and field experiments. *J. Behav. Exp. Finan.* 9, 88–97. doi: 10.1016/j.jbef.2015.12.001

Chugunova, M., and Sele, D. (2022). We and it: an interdisciplinary review of the experimental evidence on how humans interact with machines. *J. Behav. Exp. Econ.* 99, 101897. doi: 10.1016/j.socec.2022.101897

Comprehensive Taxpayer Attitude Survey (2021). *Publication 5296 (Rev. 4–2022) Catalog Number 71353Y*. US Department of the Treasury Internal Revenue Service. CTAS. Available online at: https://www.irs.gov/pub/irs-pdf/p5296.pdf (accessed July 28, 2023).

Coricelli, G., Joffily, M., Montmarquette, C., and Villeval, M. C. (2010). Cheating, emotions, and rationality: an experiment on tax evasion. *Exp. Econ.* 13, 226–247. doi: 10.1007/s10683-010-9237-5

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., et al. (2018). Cooperating with machines. *Nat. Commun.* 9, 1. doi: 10.1038/s41467-017-02597-8

Daylamani-Zad, D., and Angelides, M. C. (2021). Altruism and selfishness in believable game agents: deep reinforcement learning in modified dictator games. *IEEE Transact. Games* 13, 229–238. doi: 10.1109/TG.2020.2989636

De Melo, C., Marsella, S., and Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Trans. Comput. Hum. Interact.* 23, 8. doi: 10.1145/2890495

De Melo, C. M., Carnevale, P. J., Read, S. J., and Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *J. Pers. Soc. Psychol.* 106, 73–88. doi: 10.1037/a0034251

Dresher, M. (1962). *A Sampling Inspection Problem in Arms Control Agreements: A Game-Theoretic Analysis*. RAND Corporation. Available online at: https://www.rand.org/pubs/research_memoranda/RM2972.html (accessed July 28, 2023).

Eyssel, F., and Kuchenbrandt, D. (2012). Social categorization of social robots: anthropomorphism as a function of robot group membership. *Br. J. Soc. Psychol.* 51, 724–731. doi: 10.1111/j.2044-8309.2011.02082.x

Falk, A., and Fischbacher, U. (2002). "Crime" in the lab-detecting social interaction. *Eur. Econ. Rev.* 46, 859–869. doi: 10.1016/S0014-2921(01)00220-3

Franke, T., Attig, C., and Wessel, D. (2019). A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *Int. J. Hum. Comp. Interact.* 35, 456–467. doi: 10.1080/10447318.2018.1456150

Fritsch, S. J., Blankenheim, A., Wahl, A., Hetfeld, P., Maassen, O., Deffge, S., et al. (2022). Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. *Digit. Health* 8, 20552076221116772. doi: 10.1177/20552076221116772

Gallagher, H. L., Jack, A. I., Roepstorff, A., and Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage* 16(3, Part A), 814–821. doi: 10.1006/nimg.2002.1117

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* 19, 847–857. doi: 10.3758/s13423-012-0296-9

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

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., and Rahwan, T. (2019). Behavioural evidence for a transparency-efficiency tradeoff in human-machine cooperation. *Nat. Mach. Intell.* 1, 11. doi: 10.1038/s42256-019-0113-5

Kaber, D. B. (2018). A conceptual framework of autonomous and automated agents. *Theoret. Iss. Ergon. Sci.* 19, 406–430. doi: 10.1080/1463922X.2017.1363314

Karpus, J., Krueger, A., Verba, J. T., Bahrami, B., and Deroy, O. (2021). Algorithm exploitation: humans are keen to exploit benevolent AI. *iScience* 24, 102679. doi: 10.1016/j.isci.2021.102679

Kiesler, S., Sproull, L., and Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *J. Pers. Soc. Psychol.* 70, 47–65. doi: 10.1037/0022-3514.70.1.47

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE* 3, e2597. doi: 10.1371/journal.pone.0002597

Kushwaha, A. K., Pharswan, R., Kumar, P., and Kar, A. K. (2022). How do users feel when they use artificial intelligence for decision making? A framework for assessing users' perception. *Inf. Syst. Front.* 25, 1241–1260. doi: 10.1007/s10796-022-10293-2

Langer, E. J. (1992). Matters of mind: mindfulness/mindlessness in perspective. *Conscious. Cogn.* 1, 289–305. doi: 10.1016/1053-8100(92)90066-J

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., and 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?," in *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA), 1–28.

Lee, M., Lucas, G., and Gratch, J. (2021). Comparing mind perception in strategic exchanges: human-agent negotiation, dictator and ultimatum games. *J. Multim. User Inter.* 15, 201–214. doi: 10.1007/s12193-020-00356-6

Lefebvre, M., Pestieau, P., Riedl, A., and Villeval, M. C. (2015). Tax evasion and social information: an experiment in Belgium, France, and the Netherlands. *Int. Tax Public Finan.* 22, 401–425. doi: 10.1007/s10797-014-9318-z

Leslie, D. (2020). *Understanding Bias in Facial Recognition Technologies: An Explainer*. The Alan Turing Institute. doi: 10.2139/ssrn.3705658

Lohr, S. (2018). *Facial Recognition Is Accurate, if You're a White Guy*. New York Times. Available online at: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html (acessed July 28, 2023).

Maréchal, M., Cohn, A., and Gesche, T. (2020). *Honesty in the Digital Age*. Working Paper Series (Department of Economics). Zurich: University of Zurich.

Mascagni, G. (2018). From the lab to the field: a review of tax experiments. *J. Econ. Surv.* 32, 273–301. doi: 10.1111/joes.12201

Maschler, M. (1966). A price leadership method for solving the inspector's non-constant-sum game. *Naval Res. Logist. Q.* 13, 11–33. doi: 10.1002/nav.3800130103

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Nat. Acad. Sci. U. S. A.* 98, 11832–11835. doi: 10.1073/pnas.211415698

Nass, C., Fogg, B. J., and Moon, Y. (1996). Can computers be teammates? *Int. J. Hum. Comput. Stud.* 45, 669–678. doi: 10.1006/ijhc.1996.0073

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Nass, C., Moon, Y., and Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* 27, 864–876. doi: 10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers are social actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA), 72–78.

Nielsen, Y. A., Thielmann, I., Zettler, I., and Pfattheicher, S. (2022). Sharing money with humans versus computers: on the role of honesty-humility and (non-)social preferences. *Soc. Psychol. Personal. Sci.* 13, 1058–1068. doi: 10.1177/19485506211055622

Open AI. (2023). GPT-4 technical report. *arXiv[Preprint].arXiv: 2303.08774.* doi: 10.48550/arXiv.2303.08774

OpenAI (2022). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. Available online at: https://openai.com/blog/chatgpt/ (accessed July 28, 2023).

Palan, S., and Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *J. Behav. Exp. Finan.* 17, 22–27. doi: 10.1016/j.jbef.2017.12.004

Paolacci, G., and Chandler, J. (2014). Inside the turk: understanding mechanical turk as a participant pool. *Curr. Dir. Psychol. Sci.* 23, 184–188. doi: 10.1177/0963721414531598

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the turk: alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* 70, 153–163. doi: 10.1016/j.jesp.2017.01.006

Radu, S. (2015). "Multi-issue automated negotiation with different strategies for a car dealer business scenario," in *2015 20th International Conference on Control Systems and Computer Science*, eds I. Dumitrache, A. M. Florea, F. Pop, and A. Dumitrascu (Bucharest: IEEE), 351–356.

Rauhut, H. (2009). Higher punishment, less control? Experimental evidence on the inspection game. *Rational. Soc.* 21, 359–392. doi: 10.1177/1043463109337876

Rauhut, H. (2015). Stronger inspection incentives, less crime? Further experimental evidence on inspection games. *Ration. Soc.* 27, 414–454. doi: 10.1177/1043463115576140

Rauhut, H., and Jud, S. (2014). Avoiding detection or reciprocating norm violations? An experimental comparison of self- and other-regarding mechanisms for norm adherence. *Soz. Welt Zeitschr. Sozialwissenschaftliche Forschung Praxis* 65, 153–183. doi: 10.5771/0038-6073-2014-2-153

Rauhut, H., and Junker, M. (2009). Punishment deters crime because humans are bounded in their strategic decision-making. *J. Artif. Soc. Soc. Simul.* 12, 1. Available online at: http://jasss.soc.surrey.ac.uk/12/3/1.html

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). "Overtrust of robots in emergency evacuation scenarios," in *HRI 2016 - 11th ACM/IEEE International Conference on Human Robot Interaction* (Christchurch), 101–108.

Roose, K. (2023). *Don't ban ChatGPT in schools. Teach with It*. The New York Times. Available online at: https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html (accessed July 28, 2023).

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, OR), 141–148.

Schniter, E., Shields, T. W., and Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *J. Econ. Psychol.* 78, 102253. doi: 10.1016/j.joep.2020.102253

Selwyn, N., Cordoba, B. G., Andrejevic, M., and Campbell, L. (2020). *AI For Social Good: Australian Public Attitudes Toward AI and Society*. Monash University.

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

Spicer, M. W., and Thomas, J. E. (1982). Audit probabilities and the tax evasion decision: an experimental approach. *J. Econ. Psychol.* 2, 241–245. doi: 10.1016/0167-4870(82)90006-X

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

Torgler, B. (2002). Speaking to theorists and searching for facts: tax morale and tax compliance in experiments. *J. Econ. Surv.* 16, 657–683. doi: 10.1111/1467-6419.00185

Torgler, B. (2007). *Tax Compliance and Tax Morale*. Edward Elgar Publishing. Available online at: https://econpapers.repec.org/bookchap/elgeebook/4096.htm (accessed July 28, 2023).

Tsebelis, G. (1989). The abuse of probability in political analysis: the Robinson Crusoe fallacy. *Am. Polit. Sci. Rev.* 83, 77–91. doi: 10.2307/1956435

Tsebelis, G. (1990). Penalty has no impact on crime: a game-theoretic analysis. *Ration. Soc.* 2, 255–286. doi: 10.1177/1043463190002003002

Wang, R., Harper, F. M., and Zhu, H. (2020). "Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual differences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI), 1–14.

Wärneryd, K.-E., and Walerud, B. (1982). Taxes and economic behavior: some interview data on tax evasion in Sweden. *J. Econ. Psychol.* 2, 187–211. doi: 10.1016/0167-4870(82)90003-4

Webley, P., and Halstead, S. (1986). Tax evasion on the micro: significant simulations or expedient experiments? *J. Interdiscipl. Econ.* 1, 87–100. doi: 10.1177/02601079X8600100204

Weiss, M., Rodrigues, J., Paelecke, M., and Hewig, J. (2020). We, them, and it: dictator game offers depend on hierarchical social status, artificial intelligence, and social dominance. *Front. Psychol.* 11, 541756. doi: 10.3389/fpsyg.2020.541756

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). "Explainable AI: a brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing*, eds J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan (Dunhuang: Springer International Publishing), 563–574.

Zhang, B., and Dafoe, A. (2019). *Artificial Intelligence: American Attitudes and Trends (SSRN Scholarly Paper 3312874).*