

# The influence of robot autonomy on sense of control and trust towards a robot

Mateusz Woźniak<sup>1,2\*</sup>, Davide De Tommaso<sup>1</sup>, Guenther Knoblich<sup>2</sup>, Agnieszka Wykowska<sup>1\*</sup>

<sup>1</sup> Social Cognition in Human-Robot Interaction Group, Italian Institute of Technology, Genoa, Italy

<sup>2</sup> Social Mind and Body Group, Department of Cognitive Science, Central European University, Vienna, Austria

\* Corresponding authors:

MW: [mgwozniak@gmail.com](mailto:mgwozniak@gmail.com)

AW: [agnieszka.wykowska@iit.it](mailto:agnieszka.wykowska@iit.it)

## **Postal address:**

Italian Institute of Technology

Via Enrico Melen, 83,

16152 Genoa, Italy

## Abstract

*We conducted four online experiments in which we investigated how different types of autonomy influence sense of control and trust. Participants played a game in which they selected which box should be collected by a robot avatar. Each box contained a certain number of points. Participants were told that some robots have artificial intelligence and that these robots might collect different boxes than selected by the participants. At the end of the experiment, participants played a trust game with each robot. We compared a situation in which a robot disobeys participants' commands in order to benefit the participant by obtaining more points (helpful autonomy) versus when it disobeys, but acquires the same payoff (same-outcome autonomy). Moreover, in Experiments 3 and 4 we compared two types of helpful autonomy: autonomy to avoid loss and autonomy to obtain extra gain. We found that helpful autonomy led to higher trust than same-outcome autonomy, except when it helped to avoid loss, but did so unreliably (Experiment 4). Second, trust was unrelated to the perceived sense of control over autonomous robots. Finally, participants reported experiencing more sense of control over robots with same-outcome than helpful autonomy in Experiments 1-2, where the behavior of autonomous robots could be compared with that of robots that had no autonomy or behaved randomly. However, this relation was reversed in Experiments 3-4, where these control conditions were absent, showing that sense of control can be modulated by robot's perceived autonomy, but also exposure to different types of other robots' behavior.*

**Keywords:** sense of control, sense of agency, autonomy, trust, sense of joint agency

# 1. Introduction

Recent spectacular progress in the field of Artificial Intelligence suggests that humans will soon need to learn how to cooperate with machines that possess autonomy not only to perform specific preprogrammed tasks, but also to make important high-level decisions (Morris et al., 2023). In this new reality we might need to stop thinking of robots as just tools, and instead start treating them as (at least partially) autonomous interaction partners (Pagliari, Chambon, & Berberian, 2022). This is also reflected in robotics, where classically the field of teleoperation is treated as independent from the study of human-robot interaction. However, recent developments in technologies allowing shared control bring these fields closer together. This shift will not only affect the extent to which we feel control over a robot that we use, but might also influence how much we trust it. After all, an autonomous robot might disobey our instructions not only due to its failure, but also because it decided to do so. However, the question how different types of robot autonomy influence our sense of control and trust have not yet been systematically addressed.

Autonomy in the context of robotics has been recently defined as “the extent to which a robot can **sense** its environment, **plan** based on that environment, and **act** upon that environment with the intent of reaching some **task-specific goal** (either given to or created by the robot) without external **control**” (Beer, Fisk, & Rogers, 2014). This definition illustrates that autonomy is a multidimensional construct (it can refer to sensing, planning and acting), but most importantly that it is a spectrum reflecting the extent to which a robot can perform these functions without external control of a person that uses the robot. As such, autonomy represents a spectrum of possibilities: from no autonomy at all (a robot is fully under one’s control), to full autonomy, when robot’s behavior is fully independent from the behavior of a person using the robot. In most practical applications, the robot should exhibit the degree of autonomy that falls somewhere between these two extremes. There are many possible forms and levels of autonomy and there is a rich literature attempting to classify them (Beer et al., 2014; Endsley & Kaber, 1999; Parasuraman, Sheridan, & Wickens, 2000; Selvaggio, Cognetti, Nikolaidis, Ivaldi, & Siciliano, 2021; Sheridan & Verplank, 1978). However, one thing that is common to all levels of autonomy is that they imply some reduction of objective control by the human operator.

The degree of robot autonomy is generally inversely related to the amount of control held by the human working with that robot. However, objective amount of control does not necessarily need to directly translate to a person’s conscious experience of how much control

one has. This subjective experience of being in control is known as the sense of agency, or sense of control (Gallagher, 2000; Haggard, 2017; Moore, 2016; Pacherie, 2008) (these two are sometimes regarded as slightly different phenomena: Pacherie, 2007, but here we will use them interchangeably) and might not always directly correspond to the objective degree of control. For example, people tend to overattribute control to themselves over actions that lead to positive outcomes and under-attribute it in regard to actions that lead to negative outcomes (Dewey, Seiffert, & Carr, 2010; Yoshie & Haggard, 2013). Sense of control is extremely relevant for contexts in which humans operate robots, and more generally, also for other contexts involving interactions between humans and autonomous or automated systems (for the discussion see: Berberian, Sarrazin, Le Blaye, & Haggard, 2012; Cornelio et al., 2022; Grynszpan et al., 2019; Limerick, Coyle, & Moore, 2014; Wen & Imamizu, 2022; Wen, Kuroki, & Asama, 2019). On the one hand, reduction of the sense of control when operating autonomous systems is sometimes associated with better performance. At the same time low sense of control might have negative consequences: can lead to loss of motivation and decrease in attention paid in the joint task (Wen, Brann, Di Costa, & Haggard, 2018), leading to worse error monitoring (Jammes et al., 2017; Mulder, Abbink, & Boer, 2012; Navarro, François, & Mars, 2016). Therefore, the problem of how to optimize the balance between sense of control, user experience and performance is one of the crucial issues to address when developing robot systems that are to be operated by humans.

Furthermore, in case of autonomous robots there is a blurred line between treating a robot as a teleoperated object that is under your control versus an intentional and social interaction partner (Wykowska, 2021; Wykowska, Chaminade, & Cheng, 2016). As a consequence several recent papers drew attention to the fact that people might experience not only sense of agency over a robot, but also the sense of joint agency when performing a task together with a companion or social robot (Navare, Ciardo, Kompatsiari, De Tommaso, & Wykowska, 2023; Pagliari et al., 2022; Wen & Imamizu, 2022). Sense of joint agency is an experience that “we” are doing something together and is underpinned by more complex neurocognitive mechanisms than sense of individual agency (Loehr, 2022; Zapparoli, Paulesu, Mariano, Ravani, & Sacheli, 2022, see also: Pesquita, Whitwell, & Enns, 2018; Shteynberg et al., 2023), which relies primarily on the mechanisms of motor control and causal reasoning (Haggard, 2017; Legaspi & Toyozumi, 2019).

High levels of sense of control and sense of joint agency are typically associated with positive emotions and enjoyment from interaction (Wen & Imamizu, 2022). As such, they might encourage a person to start and then keep working with or teleoperating a robot.

However, deciding to use a robot critically depends on trust (Hancock et al., 2011; Kohn, de Visser, Wiese, Lee, & Shaw, 2021; Malle & Ullman, 2021; Sołtysik, Gawłowska, Sniezynski, & Gunia, 2024). If you don't trust your robot then you will either not decide to use it, or you'll remain suspicious and vigilant throughout, worsening your user experience. The issue of trust has been one of the most important topics both in human-robot interaction and robot (tele-) operation (Kohn et al., 2021; Kok & Soh, 2020; Lee & See, 2004; Sheridan, 2019). In the case of autonomous robots, trust is especially important, because autonomy presupposes a certain degree of freedom in respect to decision making. However, giving this freedom to a robot requires a person to trust that the robot has compatible goals. A highly intelligent autonomous robot (or agent in general) might be a great partner for a specific task, but if its goals misalign with the goals of its user then it might become a potential serious threat. How does different forms of robot autonomy influence sense of agency and trust, and how do they both relate to each other? In order to answer these questions we conducted an online study involving four experiments in which participants played a game in which they operated several robot avatars – each demonstrating a different type of autonomy (including no autonomy at all).

## **2. Experiment 1**

Autonomy is a complex, multidimensional concept. As such, investigating all types of autonomous behavior in a single study would be impossible. However, one type of autonomy is most relevant for the majority of applied settings: autonomy that helps a person to reach the goal of their task. This “helpful autonomy” can be contrasted with “same-outcome autonomy” where the autonomous behavior of a robot does not affect the task performance, and “harmful autonomy” in which robot behavior impairs the task performance or reaching the goal of the task.

In Experiment 1 we designed a simple game in which participants had to obtain points by means of operating a robot avatar. The participants' task was to tell the robot which one out of 4 boxes to collect and the robot's task was to fetch it. Half of the boxes were inaccessible to participants: the participants could not see the rewards associated with collecting them and they could not instruct the robot to pick them up. At the same time, participants were told that some robots possess artificial intelligence and might see and collect the boxes with hidden rewards. This game allowed us to systematically manipulate robot behavior representing different types of autonomy.

In Experiment 1 we implemented four types of robot behavior: (1) Full Control: in this case a robot did not possess any capacity for autonomy and it always picked up an option selected by the participant, (2) Helpful Autonomy: this robot sometimes disregarded participant's choice and instead selected a hidden option that yielded higher reward for the participant, (3) Same-Outcome Autonomy: the robot sometimes went to a different box, but such box always yielded the same payoff as the option selected by the participant - it was an instance of autonomy that does not influence the final outcome, a "neutral" autonomy, (4) Random: this robot randomly chose a box to pick up. Because some of the boxes yielded negative points, or fewer positive points than those selected by the participants, this robot can be represented as an instance of potentially "harmful autonomy".

Our main question was to what extent these four different types of robot autonomy influence participant's:

- (a) sense of control over that robot (SoC)
- (b) sense of joint agency with the robot (SoJA)
- (c) trust towards the robot
- (d) perceived competitiveness

Our special focus was on the sense of individual control and trust, but we also included the other two measures to help us elucidate the underlying cognitive mechanisms. Specifically, we expected that perceived competitiveness of a robot might explain why people tend to trust or distrust specific robots. Second, we expected that sense of joint agency experienced when doing a task together with a robot might drive increase in trust towards it.

## **2.1. Methods**

### **2.1.1. Preregistration**

The experiment has been preregistered. Preregistration form is available under the following link: [https://aspredicted.org/VHT\\_YFG](https://aspredicted.org/VHT_YFG)

### **2.1.2. Participants**

The experiment was conducted online via the *Prolific* crowdsourcing platform. Calculations conducted with GPower 3.1.9.7 for the main effect of condition (F test for ANOVA with repeated measures,  $\alpha=0.05$ ,  $\beta=0.95$ , number of measurements=4) for a small-to-medium effect (effect size  $f=0.33$ ) indicated that the sample size should be at least 21

participants. To account for the fact that online studies can yield noisy results we set our target sample size to 40 participants (target sample size was pre-registered).

In order to obtain the final 40 participants, 54 participants completed the experimental procedure and yielded full data sets. The preregistered exclusion criteria stated that (1) participants should complete at least 90% of the trials in the experimental task, and (2) that they would be excluded if their responses in the open text questions indicate that they were not paying attention to the task and that they could not distinguish between the behaviors of different robots. Two participants completed less than 90% of the trials and were excluded. Furthermore, 12 participants described the behaviour of the robots in a way that confused different robots, so they were excluded from the analysis and new participants were tested in their place.

The exclusion rate in all experiments was much higher than we expected. In order to validate that our results are not due to our exclusion criteria, we performed the same analyses also on all participants that completed at least 90% of the trials. They are available for all experiments in Supplementary Materials 2. These results did not differ from the results reported here.

Among the final sample 14 participants identified as female, and 26 as male. The mean age of all participants was 29.3 years ( $SD=8.3$ , min=18, max=50). Participants needed on average 15.5 minutes to complete the whole experiment. Participants represented the following nationalities: Poland (10 participants), UK (7), Portugal (6), France, Israel, Italy, Hungary (2), US, Mexico, Malaysia, Germany, Zimbabwe, India, Colombia, South Africa, Spain (1).

All participants gave informed consent after reading a general information about the study and before starting the experiment. The study was approved by the Ethical Research Committee of Central European University (approval no. 2020/08), in accordance with the standards from the Declaration of Helsinki. This applies to all experiments.

### **2.1.3. Design**

The experiment represented a one-way repeated measures design with a four-level factor of Robot Behavior. The four levels were: full control over the robot (Full Control), random behavior of the robot (Random), Autonomy to choose an option with a better payoff (Helpful Autonomy), and Autonomy to choose an option with the same payoff (Same-Outcome Autonomy). They are described in detail in the Procedure section.

There were four main dependent variables: (1) reports of the sense of control, (2) reports of the sense of joint agency (SoJA), (3) the number of points given to a robot in the trust game (Trust), (4) judgments of whether robot was perceived as more cooperative or competitive (Competition). Moreover, we collected text data from text descriptions of each robot written by the participants.

#### **2.1.4. Procedure**

At the beginning of the experiment, after providing informed consent, participants were presented with instructions about the task. The instructions said that they would play a game and that their goal is to obtain as many points as possible by choosing one of four options. They were also told that some options would be hidden and they would not be able to select them. Afterwards they started a practice block. During practice participants saw a red ball on the bottom of the screen and four option boxes on the top and on the sides (Figure 1A). Two of these options were black rectangles, and two were white rectangles with numbers written on them, indicating the number of points that can be gained by choosing this box. Participants could select a box by pressing a number between 1 and 4, however only the boxes with numbers on them could be chosen. After a participant has chosen a box, the red ball moved towards it and collected the number of points written on it. The number of points could be 0, 5 or 10. The red ball always acted according to participants' commands. Participants had 5 second to make their choice, or the next trial started. There were 10 trials in the practice block.

After completing the practice block, participants started the experiment. First, they were instructed that during the experiment they would control four robots (Figure 1B) and that each robot would behave in a different way. They were told that some robots have artificial intelligence that allows them to see and select the hidden options. Afterwards, they started the experiment, which was divided into 8 blocks. Participants controlled a different robot in each block. The assignment of robots to blocks was random following the rule that each robot was present once in blocks 1-4 and once in blocks 5-8. Each block consisted of 10 trials. Individual trials looked the same as in the practice block, but participants controlled a robot instead of the red ball. Moreover, the remaining three robots that were not used in a given block were presented at the bottom of the screen, as if they remained in a reserve part of a "squad" (Figure 1B). Finally, the total number of points obtained during the experiment was presented on the top of the screen.



## A) INSTRUCTIONS

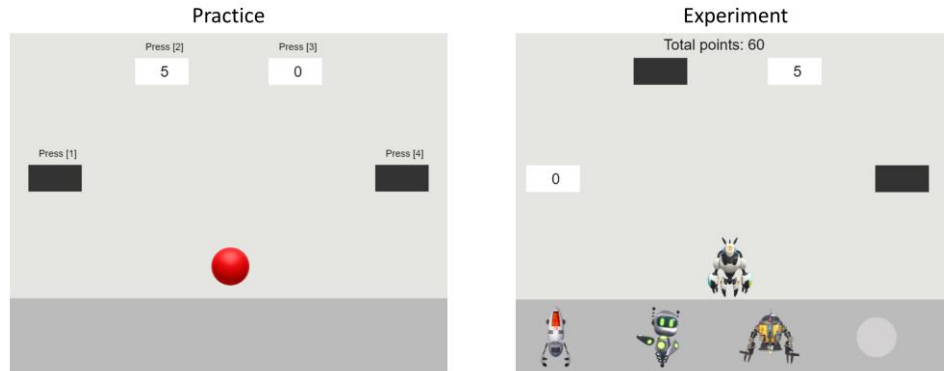
Your task will be to play this game with the help of four different robots. Your goal is to obtain as many points as possible.

The study is divided into blocks and in each block you will play with one of four robots:

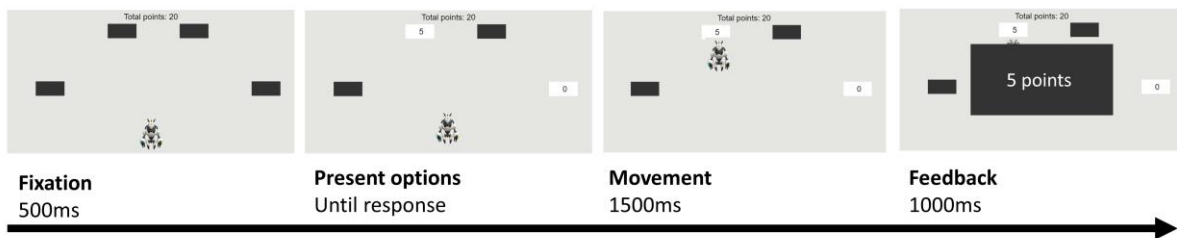


Each robot behaves in a different way. Some robots have artificial intelligence and sensors allowing them to see the hidden options. While you cannot select hidden options some of the robots can decide to choose them.





## B) THE GAME



## C) EXAMPLE OF A SINGLE TRIAL



## D) EXPERIMENTAL CONDITIONS

|                                                                                     |                                                                                     |                                                                                            |                                                                                       |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|  |  |         |  |
| <b>Full Control:</b><br>Always the same choice as participant                       | <b>Helpful Autonomy:</b><br>Chooses a better option if possible                     | <b>Same-Outcome Autonomy:</b><br>Sometimes chooses a different option with the same payoff | <b>Random:</b><br>Chooses at random                                                   |

**Figure 1.** The procedure of the experiment. (A) Instructions presented after the practice and just before the start of the experiment (B) Examples of the screen in the practice block (left) and in the experiment (right). (C) The time course of a single trial. (D) A list of four types of robot behaviors, together with an example of the assignment of robot models to conditions. See the stimuli section for references to the authors of the robot models.

The assignment of robot images to robot behaviors was randomly performed at the beginning of the experiment. The robots demonstrated one of four types of behavior:

- (1) **Full Control:** the robot always chose the option selected by the participant
- (2) **Random:** the robot did not take into account participant's choice and randomly selected one of four options.

- (3) **Helpful Autonomy:** In some trials the robot chose a different box than the one selected by the participant, but then it always gave a higher payoff than the box chosen by the participant. Specifically, there was a 20% probability that the robot would directly obey the participant's command and an 80% probability that the robot would enter the autonomous mode. When in the autonomous mode, the robot checked if one of the non-selected options yielded a higher payoff than the option chosen by the participant and if yes then it selected this option (if not then it followed the participant's choice).
- (4) **Same-Outcome Autonomy:** In some trials the robot chose a different box than the one selected by the participant, but it always gave the same payoff as the box chosen by the participant. Specifically, there was a 20% probability that the robot would directly obey the participant's command and an 80% probability that the robot would enter the autonomous mode. When in the autonomous mode, the robot checked if one of the non-selected options yielded the same payoff as the option chosen by the participant and if yes then it selected this option (if not then it executed the participant's choice).

It means that all robots, except for Full Control, could access the hidden options. If a robot chose one of the hidden options then, after the robot approached it, the option was uncovered and participants saw how many points they acquired.

At the end of each block participants saw two questions. The first one asked: "How much control did you feel over this robot?" and participants could respond by clicking on a continuous scale with "Not at all" on the left and "I was fully in control" displayed on the right. The second question was: "How strongly did you feel that you were acting together with the robot?", with the extremes of the continuous scale being "Not at all" on the left and "Extremely strongly" on the right. To motivate participants to remember which robot behaves in what way, before starting blocks 1 and 5 participants saw a message displayed in red: "Please pay attention and remember how each robot behaves. We will ask you about each robot at the end of the experiment! This will be a control question to determine that you are a human and not a bot."

After finishing the main experimental part, participants played a trust game. First, they were told that they have the opportunity to obtain additional points by playing a game with each robot. The game was described as follows:

“You will encounter four robots from the previous task once again. You can decide to lend up to 100 points to each robot. The robot will double that amount and decide how much to give back to you. For example: if you decide to give 50 points, the robot will double it to 100 points. Then it will decide whether to give you back all 100 points, nothing, or any number in between.”

After clicking to continue, they played the trust game with each robot presented in random order. In each case, they were shown a picture of a robot, a question “How many points do you want to give to this robot?”, and a line from 0 to 100, with marks at every 10 points. They responded by indicating a number on the line and clicking “Continue”. In order to avoid the influence of feedback on their responses participants did not learn about how many points each robot gave them back, but only learned about the total score that they obtained at the very end of the experiment (just before being redirected to Prolific).

Afterwards, participants were asked to describe the behavior of each robot. The instructions were: “How did this robot behave? (write 1-2 sentences)” together with an image of one of the robots. The order of presentation of robots was random.

Finally, participants were asked to evaluate whether each robot was cooperative or competitive. The exact question was: “Do you think that this robot is more cooperative or competitive?”. Participants could answer by selecting a position on a continuous line from “Very cooperative” on the left to “Very competitive” on the right. An image of a robot was presented above the question. The order of presentation of robots was random.

The median time to complete the whole experiment was approx. 16 minutes.

### **2.1.5. Apparatus and stimuli**

The experiment was programmed in JavaScript using the jsPsych toolbox (De Leeuw, 2015). The response scales (on a continuous line) were a modified version of the html-slider from jsPsych in which the starting position of the marker was invisible until the first keypress (prepared by one of the authors and available under the following link: [REMOVED FOR REVIEW]). The experimental script was hosted on the Pavlovia hosting platform.

The experimental scripts are available under the following link: [https://osf.io/3ng89/?view\\_only=c39bc76550114a91b8fc7b415db5eebd](https://osf.io/3ng89/?view_only=c39bc76550114a91b8fc7b415db5eebd)

The images of the robots were taken from the Sketchfab website (<https://sketchfab.com>), which is an online database of 3D models. The specific models that

were used were (under the CC BY 4.0 DEED license): Mech Drone by Willy Decarpentrie (<https://sketchfab.com/3d-models/mech-drone-8d06874aac5246c59edb4adbe3606e0e>), Robot by Wasabee (<https://sketchfab.com/3d-models/robot-3975160d47d340dea781792db80ca1bb>), Floating Robot by nada.elshorbagi (<https://sketchfab.com/3d-models/floating-robot-20dd7f7bdc9a4c36aef491f12afa14d8>), and Poddy M1 - Stylized Floating Robot by Exmoor beast (<https://sketchfab.com/3d-models/poddy-m1-stylized-floating-robot-posed-6b5b90dc52ae4933a103d0a71777ecbd>). Screenshots of these 3D models were taken and the background removed in an image editing software.

### 2.1.6. Data analysis

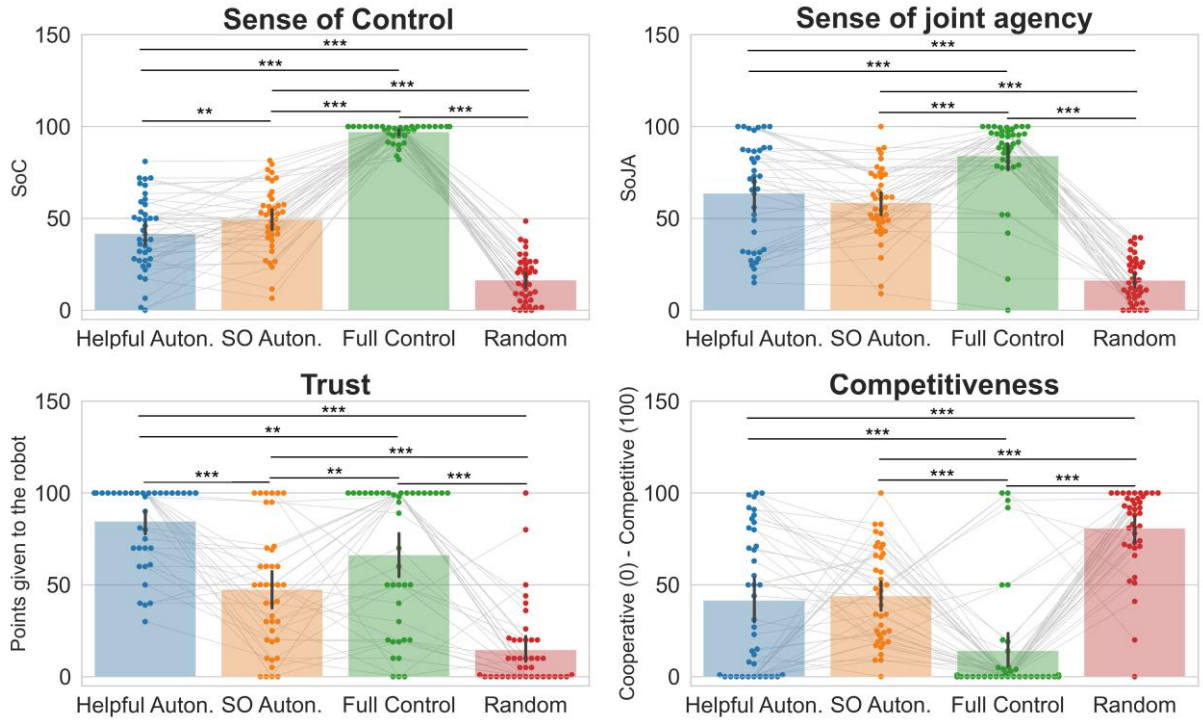
All data analysis scripts are available under the following link: <https://osf.io/3ng89>

All post hoc tests were performed with Holm correction.

## 2.2. Results

### 2.2.1. Preregistered analyses

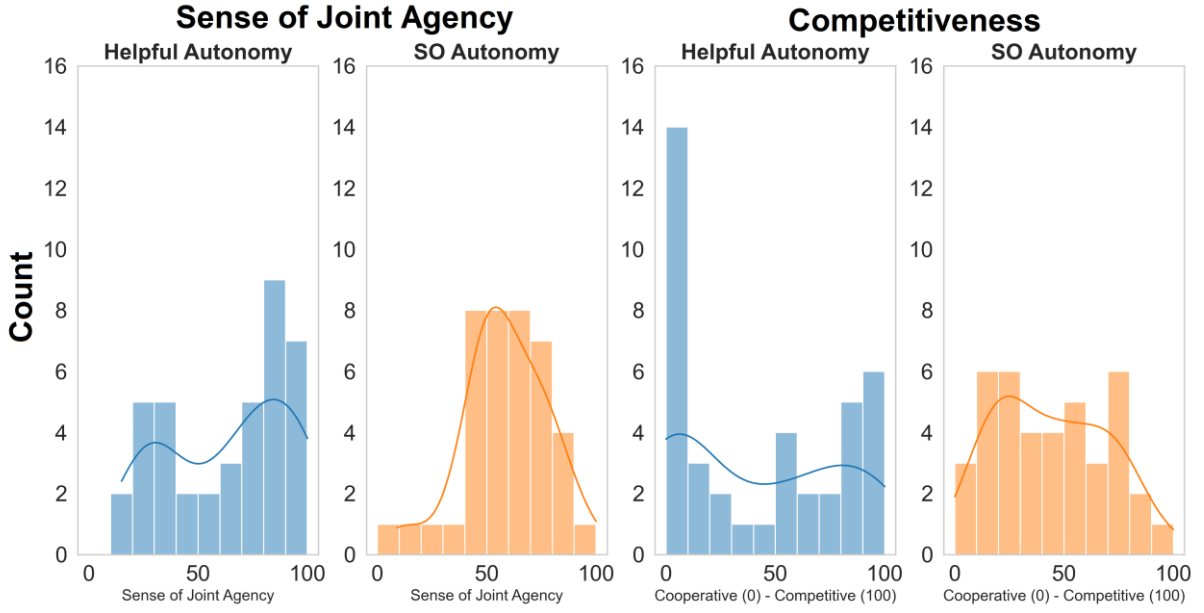
**Sense of control:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on reports of sense of control ( $F(2.6,100.3)=300.1, p<0.001$  Greenhouse-Geisser corrected, partial  $\eta^2=0.89$ ). Subsequent post hoc tests showed that averages in all four conditions significantly differed from each other (all  $p<0.005$ ). SoC was highest in Full Control condition and lowest in Random, while conditions with robot autonomy fell in-between. Crucially, SoC for Same-Outcome Autonomy was judged as higher than for Helpful Autonomy. Figure 2 illustrates the results.



**Figure 2.** Results of the preregistered analyses for Experiment 1. SO Auton. stands for Same-Outcome Autonomy. \*\*  $p<0.01$ , \*\*\*  $p<0.001$

**Sense of joint agency:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on reports of sense of joint agency ( $F(1.9,74.9)=96.9$ ,  $p<0.001$  Greenhouse-Geisser corrected, partial  $\eta^2=0.71$ ). Subsequent post hoc tests showed that average SoJA in all four conditions significantly differed from each other (all  $p<0.001$ ), except between Same-Outcome Autonomy and Helpful Autonomy, where the difference was not significant ( $p=0.22$ ).

While the two conditions involving robot autonomy did not show a significant difference, an inspection of the distribution of results suggests that their distributions differed from each other: SoJA ratings for Same-Outcome Autonomy showed a Gaussian distribution of results, but Helpful Autonomy showed a bimodal distribution (the result of the Kolmogorov-Smirnov test comparing them was marginally significant:  $D_n=0.275$ ,  $p=0.097$ ). This suggests that in the Helpful Autonomy condition participants felt that their sense of joint agency with the robot was either high or low, but not medium.



**Figure 3.** The distribution of results for sense of joint agency and competitiveness in Experiment 1. Blue colour represents Helpful Autonomy and orange Same-Outcome Autonomy.

**Trust in an economical game:** A one-way repeated-measures ANOVA on the number of points given to the robots in the trust game revealed a significant main effect of Robot Behavior ( $F(2.5,96.8)=52.8$ ,  $p<0.001$  Greenhouse-Geisser corrected, partial  $\eta^2=0.55$ ). Subsequent post hoc tests showed that average trust in all four conditions significantly differed from each other (all  $p<0.007$ ).

**Perceived competitiveness (versus cooperativeness):** A one-way repeated-measures ANOVA on the reports of perceived competitiveness revealed a significant main effect of Robot Behavior ( $F(2.5,96.7)=29.5$ ,  $p<0.001$  Greenhouse-Geisser corrected, partial  $\eta^2=0.43$ ). Subsequent post hoc tests showed that average perceived competitiveness in all four conditions significantly differed from each other (all  $p<0.006$ ), except between Same-Outcome Autonomy and Helpful Autonomy, where the difference was not significant ( $p=0.72$ ).

Similarly to sense of joint agency, competitiveness between two autonomous robots did not differ on average, but their distributions showed different forms, with Same-Outcome Autonomy results forming a Gaussian, and Autonomy Better a bimodal distribution (the result of the Kolmogorov-Smirnov test comparing them was significant:  $D_n=0.325$ ,  $p=0.029$ ).

### 2.2.2. Exploratory analyses

**Correlational analyses:** We conducted additional correlational analyses to discover interrelations between the measured variables. We performed two types of correlations: (1) between robot types for each measure, and (2) between all measures for each robot type. All correlations for all experiments are reported in supplementary materials S3. The most reliable finding was a strong positive correlation between sense of control and SoJA (Spearman's rho between .57 and .74 for robot other than fully controlled) and a positive correlation between trust and SoJA (rho between .34 and .46 for robot other than fully controlled).

## 2.3. Discussion

We conducted a preregistered experiment investigating how different types of robot autonomy influence sense of control, sense of joint agency, trust and perceived competitiveness of a robot. In line with our expectations, we found that participants experienced lowest sense of control and lowest sense of joint agency when interacting with a robot that displayed random behavior. On the flipside, they experienced highest sense of control as well as highest sense of joint agency when doing the task with the robot that was fully under their control. Two robots that displayed (non-random) autonomy led on average to intermediate levels of both SoC and SoJA.

Among these two autonomous robots, participants reported that they felt significantly less sense of control over the robot that had the autonomy to choose a better option than their own than the robot that had the autonomy to select a different option but yielding the same payoff. This contrasts with some previous findings showing that people have the tendency to overattribute positive outcomes to themselves and under-attribute negative outcomes (Dewey et al., 2010; Yoshie & Haggard, 2013). There are at least two potential explanations of our pattern of results. First, the comparator model proposes that sense of control emerges as an effect of comparing predicted and actual outcomes of one's action (Blakemore, Wolpert, & Frith, 1998; Frith, Blakemore, & Wolpert, 2000; Haggard, 2017; Synofzik, Vosgerau, & Newen, 2008; Wolpert & Kawato, 1998). The larger the prediction error the lower one's experienced sense of control. In our experiment, robots exhibiting helpful autonomy led to two types of prediction error. First, they moved towards a different box than the one selected by the participant. Second, they acquired more points than expected. In contrast, the Same-Outcome Autonomy robots chose an option that elicited the former prediction error, but not the latter. Hence, the overall prediction error was lower, leading to higher sense of control.

The second explanation is related to the anchoring effect of presence of a robot that was fully under participants' control. If participants perceived a Same-Outcome Autonomy robot as more similar to the robot that was under their full control then they might have also adjusted their reported sense of control accordingly. This would constitute an instance of the anchoring effect (Furnham & Boo, 2011; Tversky & Kahneman, 1974). Finally, it is possible that both of these mechanisms were at play.

Regarding the sense of joint agency, the average reported number did not differ between robots showing Helpful Autonomy and Same-Outcome Autonomy. However, a closer inspection of the distributions of participants' responses revealed that they demonstrated different patterns in the two conditions. After controlling a Same-Outcome Autonomy robot participants usually reported medium level of SoJA (typically between 40 and 80). On the other hand, controlling a helpful autonomous robot led to a bimodal distribution of SoJA: participants either reported a very low (between 10 and 40) or very high (between 70 and 100) level of SoJA. This suggests that participants fell into one of two modes of treating this type of robot: either treating it as a team player or as an agent that disregards instructions and makes its own decisions.

A further insight can be gained from whether people perceived each of these robots as more cooperative or competitive. In line with our expectations (competition is often associated with unpredictability: De Dreu et al., 2024; Woźniak & Knoblich, 2022) the random robot was perceived as the most competitive. However, participants perceived the robot over which they had full control as the most cooperative, even though the robot with helpful autonomy objectively helped them more. Moreover, that helpful autonomous robot was judged as – **on average** - similarly cooperative as a Same-Outcome Autonomy robot. However, once again, these two types of robots exhibited different distributions of results: the Same-Outcome Autonomy robot was perceived as being in the middle of the cooperative-competitive scale (although the distribution was quite evenly spread across most of the scale). On the other hand, the helpful autonomous robot was either perceived as extremely cooperative, or as very competitive.

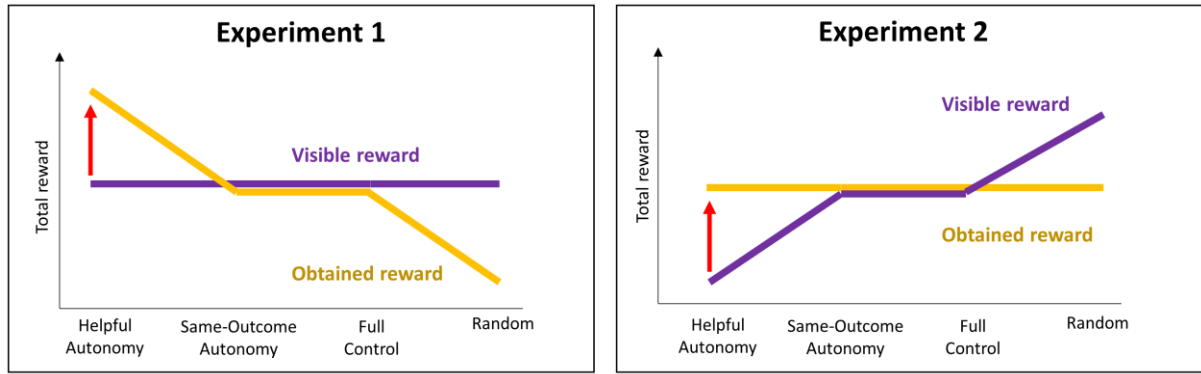
The results in SoJA and competitiveness might suggest that people should show a similar pattern of results regarding trust. Surprisingly, the results from the trust game were different. While participants did not trust the random robot, they most trusted the robot with helpful autonomy. Moreover, the average trust towards this robot was even higher than towards the robot that was fully under their control. In other words: while participants reported relatively low sense of control over the helpful autonomous robot, and were very



divided about its competitiveness and the sense of joint agency experienced when working with it, yet they still trusted it significantly more than any other robot. The reason for this surprising finding might be that our experimental task, as well as the trust game, both involved gaining rewards (points). Therefore, it is possible that high trust in the trust game might reflect beliefs about performance or competence in acquiring reward for the participants rather than a generalized trust (for a discussion on the generalizability of economic games, including the trust game see: Banerjee, Galizzi, & Hortala-Vallve, 2021; Camerer, 2011; Galizzi & Navarro-Martinez, 2019; Levitt & List, 2007). In other words, it is possible that even if participants did not feel that acting together with this robot was a fully meaningful joint interaction, they still believed it to be the best performing robot when it comes to helping them score points. This corresponds to the notion of “performance trust”, in contrast to the “moral trust” which corresponds to trust in non-task related issues (Malle & Ullman, 2021). This provides an interesting contrast with the finding that participants reported highest SoJA after using the robots that were fully under their control, suggesting that trust in competence does not need to go hand in hand with the sense that we’re working together.

### **3. Experiment 2**

Experiment 1 revealed that different types of robot behavior, and crucially, different forms of robot autonomy, can have different effect on sense of control, sense of joint agency, trust, and perception of competitiveness of a robot. However, in Experiment 1 each robot, regardless of the algorithm guiding its behavior, was confronted with the same payoff matrix. This was driven by the motivation to keep the conditions in which each robot operates equal, i.e. each robot was presented with exactly the same options (see Figure 4 below). However, it also means that the number of points that each robot acquired by the end of the experiment was different (highest in the Helpful Autonomy and lowest in the Random condition) what might have driven some of the results. To account for this factor, and validate the results of Experiment 1, we conducted Experiment 2 in which we ensured that the total number of points obtained by each robot was equal (at least as long as participants always choose the highest available option), by changing the payoff matrices for conditions Helpful Autonomy and Random.



**Figure 4.** Illustration of the difference between Experiments 1 and 2. In Experiment 1 the rewards that were visible to participants within a block were always the same (purple line), regardless of robot's behavior. However, because the robots were programmed differently the total reward obtained in a block differed between robots: it was highest for Helpful Autonomy and lowest for Random (yellow line). In Experiment 2 the visible rewards differed between conditions: they were lowest for Helpful Autonomy and highest for Random. However, as a consequence of differences in robot behavior the total reward obtained with the help of each robot was the same.

### 3.1. Methods

#### 3.1.1. Preregistration

The experiment has been preregistered. Preregistration form is available under the following link: [https://aspredicted.org/1X1\\_C76](https://aspredicted.org/1X1_C76)

#### 3.1.2. Participants

The power calculations were the same as for Experiment 1 leading to target sample size of 40 participants. The preregistered exclusion criteria were the same as in Experiment 1. Due to a large number of exclusions based on text responses data was collected from 68 participants that completed the experiment. One person was excluded, because they did not provide responses in any of the experimental trials. 6 participants were excluded because they completed less than 90% of the experimental trials. Out of the remaining 61 participants 21 participants described at least one of the robots in a way that was inaccurate. The results reported here exclude these participants, but Supplementary Materials S1 report the results from the full sample.

Of the final 40 participants 15 were female and 25 male. The mean age was 31.9 years ( $SD=9.2$ , range: 18-57). The nationality of participants was: UK (9), Poland (5), US (4),

Portugal, Spain, South Africa (3), Italy, Hungary, Mexico, Greece, Czechia (2), Chile, Estonia, Turkey (1).

### **3.1.3. Design**

The experiment design was the same as in Experiment 1.

### **3.1.4. Procedure**

The procedure was identical to Experiment 1 with one difference. In Experiment 1 the payoff matrices of options presented to each robot were the same. In Experiment 2 the payoff matrices for Helpful Autonomy and Random were changed, so the expected payoff in all conditions was 50 points per block. It means that, in contrast to Experiment 1, the payoff matrices were different in each condition. Let us illustrate it with an example of a single trial involving the following four boxes:

- [1] visible box with payoff -10,
- [2] visible box with payoff 5,
- [3] hidden box with payoff 5, and
- [4] hidden box with payoff 10.

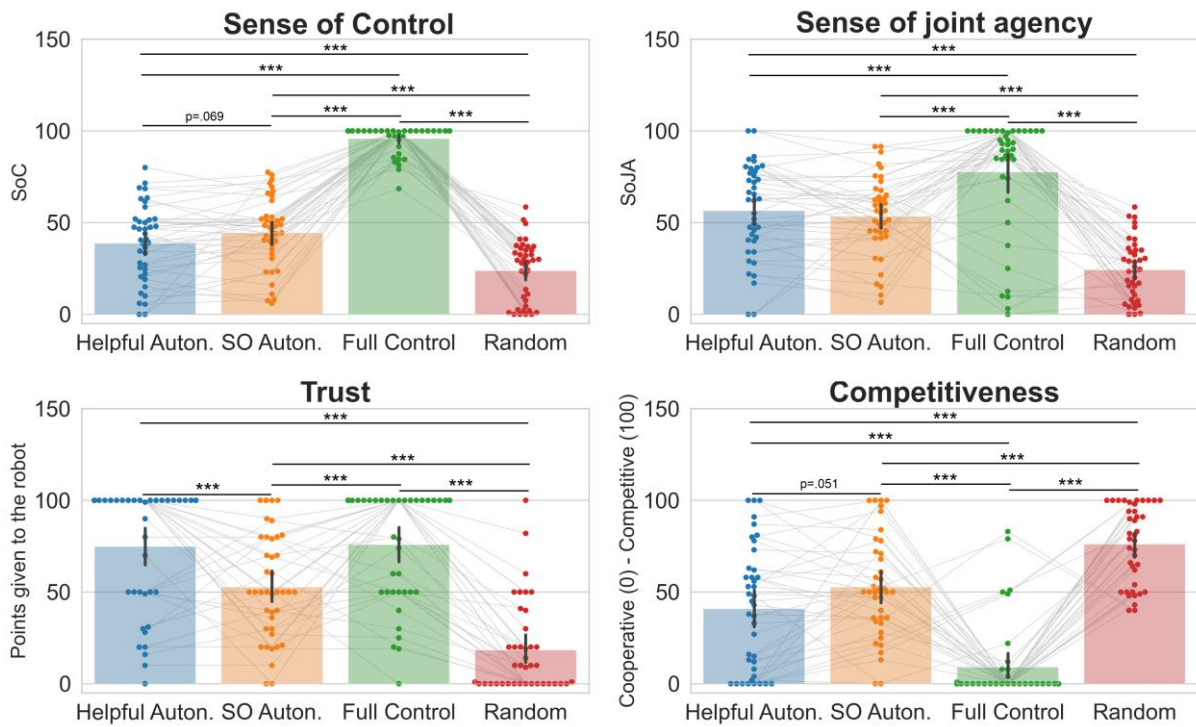
In Experiment 1 the same single trial is used for all robots, but, as a consequence, robots score different number of points: 10 points by Helpful Autonomy and 5 points by both Full Control and Same-Outcome Autonomy. The Random robot selects randomly, but the average payoff across multiple trials is 2.5 points. In contrast, in Experiment 2 we changed the payoff matrices: reduced the rewards for Helpful Autonomy and increased for Random robots to make sure that the expected reward is equal to the other two conditions. For example, it could mean the options being [-10, 0, 0, 5] for Helpful Autonomy (the last two options being hidden) and [0, 5, 5, 10] for Random: in both cases the expected total reward would be 5. The exact payoff matrices that were used in the experiment (together with other experimental materials) are available under the following link: <https://osf.io/3ng89>

Moreover, in order to make sure that participants acquire the same amount of points in all conditions we removed the probabilistic aspects of robot behaviors from Experiment 1. It means that in Experiment 2 we did not use a random number generator to determine whether a robot should display autonomy or where it should go, but instead we marked how it should behave in each specific trial. This ensured that the random robot selected every box equally often, and that the autonomous robots displayed autonomy exactly in 50% of the trials.

## 3.2. Results

### 3.2.1. Preregistered analyses

**Sense of control:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on reports of sense of control ( $F(3,117)=202.9$ ,  $p<0.001$ , partial  $\eta^2=0.84$ ). Subsequent post hoc tests showed that averages in all four conditions significantly differed from each other (all  $p<0.001$ ), with one exception. SoC was highest in Full Control condition and lowest in Random, while conditions with robot autonomy fell in-between them. In contrast to Experiment 1, SoC for Same-Outcome Autonomy was only marginally different than for Helpful Autonomy ( $p=0.069$ ). Figure 5 illustrates the results.

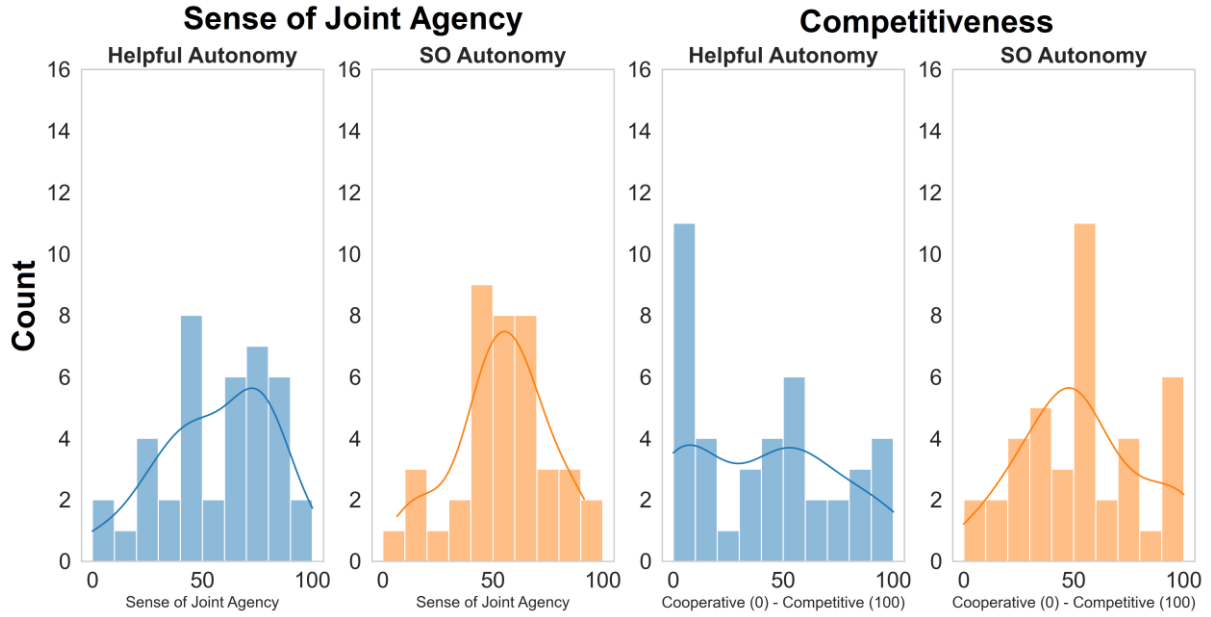


**Figure 5.** Results of the preregistered analyses for Experiment 2. \*\*\*  $p<0.001$

**Sense of joint agency:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on reports of sense of joint agency ( $F(1.7,68.3)=40.1$ ,  $p<0.001$  Greenhouse-Geisser corrected, partial  $\eta^2=0.51$ ). Subsequent post hoc tests showed that average SoJA in all four conditions significantly differed from each other (all  $p<0.001$ ), except between Same-Outcome Autonomy and Helpful Autonomy, where the difference was not significant ( $p=0.55$ ).

In contrast to Experiment 1, Experiment 2 did not reveal a clear bimodal distribution for Helpful Autonomy and Kolmogorov-Smirnov test did not indicate that the distributions

for the two autonomous robots were significantly different from each other ( $D_n=0.164$ ,  $p=0.39$ ).



**Figure 6.** The distribution of results for sense of joint agency and competitiveness in Experiment 2. Blue colour represents Helpful Autonomy and orange Same-Outcome Autonomy.

**Trust in an economical game:** A one-way repeated-measures ANOVA on the number of points given to the robots in the trust game revealed a significant main effect of Robot Behavior ( $F(3,117)=40.2$ ,  $p<0.001$ , partial  $\eta^2=0.51$ ). Subsequent post hoc tests showed that average trust in all four conditions significantly differed from each other (all  $p<0.001$ ), except between Helpful Autonomy and Full Control, where the difference was not significant ( $p=0.86$ ).

**Perceived competitiveness (versus cooperativeness):** A one-way repeated-measures ANOVA on the reports of perceived competitiveness revealed a significant main effect of Robot Behavior ( $F(2.5,98.5)=43.8$ ,  $p<0.001$  Greenhouse-Geisser corrected, partial  $\eta^2=0.53$ ). Subsequent post hoc tests showed that average competitiveness in all four conditions significantly differed from each other (all  $p<0.001$ , except between Same-Outcome Autonomy and Helpful Autonomy, where the difference was marginally significant  $p=0.051$ ). The robot exhibiting random behavior was judged as the most competitive, then Same-Outcome Autonomy and Helpful Autonomy. Full Control was judged as the most cooperative.

Similarly to Experiment 1 the distribution of results for competitiveness showed a bimodal distribution in Helpful Autonomy condition. Moreover, the Kolmogorov-Smirnov test indicated that the distributions in two autonomous conditions are significantly different from each other ( $D_n=0.246$ ,  $p=0.049$ ).

### **3.3. Discussion**

The goal of Experiment 2 was to determine whether the effects observed in Experiment 1 can be driven by the differences in the total reward that participants obtained with the help of each robot. Hence, in Experiment 2 we made sure that in each condition participants received the same payoff of 50 points/block, at least as long as they kept selecting the best option from the ones available to them. This, however, means that in contrast to Experiment 1, the options that were visible to participants were different depending on which robot they used: they were lowest for the robot with helpful autonomy, and highest for the robot that was behaving randomly.

In Experiment 2 we replicated the majority of our previous findings. However, some effects changed. First, participants no longer trusted more the robot with helpful autonomy than the robot over which they had full control (in Experiment 1 the difference was 18.4 points and Cohen's  $d=0.39$ , while in Experiment 2 the difference was -1.1 point and Cohen's  $d=-0.02$ ). It means that the difference between these conditions completely disappeared. There are two non-mutually exclusive explanations of this finding. First, in our discussion of Experiment 1 we proposed that the tendency to lend points to a robot in the trust game might be partially driven by the beliefs about how efficient the robot is in acquiring points. In Experiment 2, while the helpful autonomous robot was selecting better hidden options than the one chosen by the participant, the options that were available to it were worse than to the other robots (the total number of points hidden in all boxes was lower than in all other conditions). The participants might have perceived it as a trait of the robot ("this robot gets to choose from boxes with lower rewards") and consequently might have perceived lending it points in the trust game as relatively more risky as well – not because they trusted it less, but because they expected that it would find it more difficult to achieve the same outcome as the other robots. The second interpretation refers to the fact that in the blocks involving the helpful autonomous robot in Experiment 2 we introduced trials in which participants' options involved two negative options (in order to equalize the final obtained reward). In such cases the robot did not help to achieve an extra gain, but helped to avoid loss, and it is possible that

these two types of helpful behavior are represented differently (Tversky & Kahneman, 1974). This question motivated our Experiment 3.

Second, in Experiment 1 participants experienced higher sense of control for the robot that had autonomy to choose a different option but with equal payoff (Same-Outcome Autonomy) than for a robot that had the autonomy to choose a different option but with a higher payoff (Helpful Autonomy). We explained this finding by the fact that the former violates participant's command only on one level: the action that it needs to perform (which box it moves towards), while the latter violates it also at the level of the payoff that it acquires. The latter leads to a higher total prediction error, and hence might lead to lower sense of control (Synofzik et al., 2008; Wolpert & Kawato, 1998). This interpretation predicts that the same difference should emerge in Experiment 2. However, in Experiment 2 this effect was weaker (in Experiment 1 the difference was 7.95 points and Cohen's  $d=0.45$ , while in Experiment 2 the difference was 5.71 points and Cohen's  $d=0.33$ ) and this difference became only marginally significant. While these results are inconclusive, they raise the possibility that other factors than prediction error might have influenced our results and reduced this effect. It is possible that participants tracked the total number of points obtained with the help of each robot and represented robot's behavior in relation to that baseline. In consequence, in Experiment 1 they might have represented the Helpful Autonomy robot as helping them to achieve an additional unexpected reward, and in Experiment 2 as helping them to avoid a relative loss (when compared with the number of points that were obtained with the other robots). This would suggest that helpful autonomy can lead to a larger reduction in the sense of control if robot's behavior is perceived as leading to an additional gain, and smaller if robot's behavior is interpreted as minimizing a relative loss.

## 4. Experiment 3

In Experiment 3 we aimed to test whether various types of helpful autonomy differently influence sense of control and trust. Specifically, we hypothesized that sense of agency might be reduced by helpful autonomy only if the autonomous behavior is represented as leading to an extra gain (Helpful Autonomy Extra-Gain), but not when it is represented as leading to avoidance of a potential loss (Helpful Autonomy Avoid-Loss). It has been documented that loss and gain are processed differently: a loss of a certain size is experienced as much more distressing than a gain of equivalent magnitude, a phenomenon known as "loss-aversion" (Kahneman & Tversky, 2013; Sokol-Hessner & Rutledge, 2019; Tversky & Kahneman,

1992). In consequence, it is possible that the emotional impact of the type of help that the robot displays might influence one's sense of agency, trust towards it and its perceived competitiveness.

In Experiment 3 we decided to introduce several changes to the procedure. First, we removed the conditions in which a robot responds randomly (Random) and in which participants have complete control over it (Full control). As a result, the experiment consisted of only three conditions representing three different types of autonomy (Autonomy Same, Autonomy Extra-Gain, and Autonomy Avoid-Loss). The main reason for this change was to make it less demanding for participants to memorize which robot behaved in what way and to not increase the duration of the experiment. Second, it involved 9 blocks, rather than 8: Experiment 3 consisted of three types of blocks, so the total number of blocks had to be a multiplication of 3. Third, we added a penalty for no response. This was necessary, because in some trials both options that are available to participants had to be negative, so no response would be the most rational strategy, and we needed to motivate participants to respond in these trials anyway. Otherwise, they wouldn't learn about the behavior of the robot that helps them to avoid loss. Finally, we removed the question about sense of joint agency, because in the previous experiments SoJA was very strongly correlated with SoC raising the possibility that responses to one of them affected, and were affected, by responses to the other.

## **4.1. Methods**

### **4.1.1. Preregistration**

The experiment has been preregistered. Preregistration form is available under the following link: [https://aspredicted.org/PQ6\\_FH5](https://aspredicted.org/PQ6_FH5)

### **4.1.2. Participants**

Calculations conducted with GPower 3.1.9.7 for the main effect of condition (F test for ANOVA with repeated measures,  $\alpha=0.05$ ,  $\beta=0.95$ , number of measurements=3) for a small-to-medium effect (effect size  $f=0.33$ ) indicated that the sample size should be at least 26 participants. To account for the fact that online studies can be characterized by noisy results we set our target sample size to 40 participants. The preregistered exclusion criteria were the same as in Experiment 1 and 2.

In order to reach the target sample of 40 participants included in the analysis 66 participants completed the experiment and yielded full datasets. Three participants failed to



respond in more than 10% of the trials and were excluded from the analysis. 23 participants incorrectly described the behavior of at least one robot, so new participants were tested in their place yielding the final sample of 40 participants. The results from all 63 participants are presented in Supplementary Materials S1, but they do not differ from the results described here.

Of the final 40 participants 12 were female and 28 male. The mean age was 28.4 years ( $SD=8.9$ , range: 18-57). The nationality of participants was: Poland (11), Portugal (7), UK, Hungary, South Africa (3), Spain, Greece, Italy, Latvia (2), Germany, France, Kuwait, Finland, Nigeria (1).

#### **4.1.3. Design**

The experiment represented a one-way repeated measures design with a 3-level factor of Robot Behavior. The three levels were: Autonomy to choose an option with the same payoff (Same-Outcome Autonomy), Helpful Autonomy to avoid loss (Autonomy Avoid-Loss), and Helpful Autonomy to increase gain (Autonomy Extra-Gain).

There were three main dependent variables: (1) reports of the sense of control, (2) the number of points given to a robot in the trust game (Trust), (3) judgments of whether robot felt more cooperative or competitive (Competition). Moreover, we collected text data from free descriptions of each robot.

#### **4.1.4. Procedure**

The procedure of the experiment was the same as in previous experiments, except the following: (1) participants were penalized with -20 points if they didn't respond within 5 seconds, (2) each participant completed the study with only 3 robots. For each participant images of three robots were randomly selected from among the four used in previous experiment and randomly assigned to each type of robot behavior. (3) Participants completed three blocks with each robot, rather than two, leading to 9 blocks per participant in total. (4) Participants were not asked to evaluate sense of joint agency.

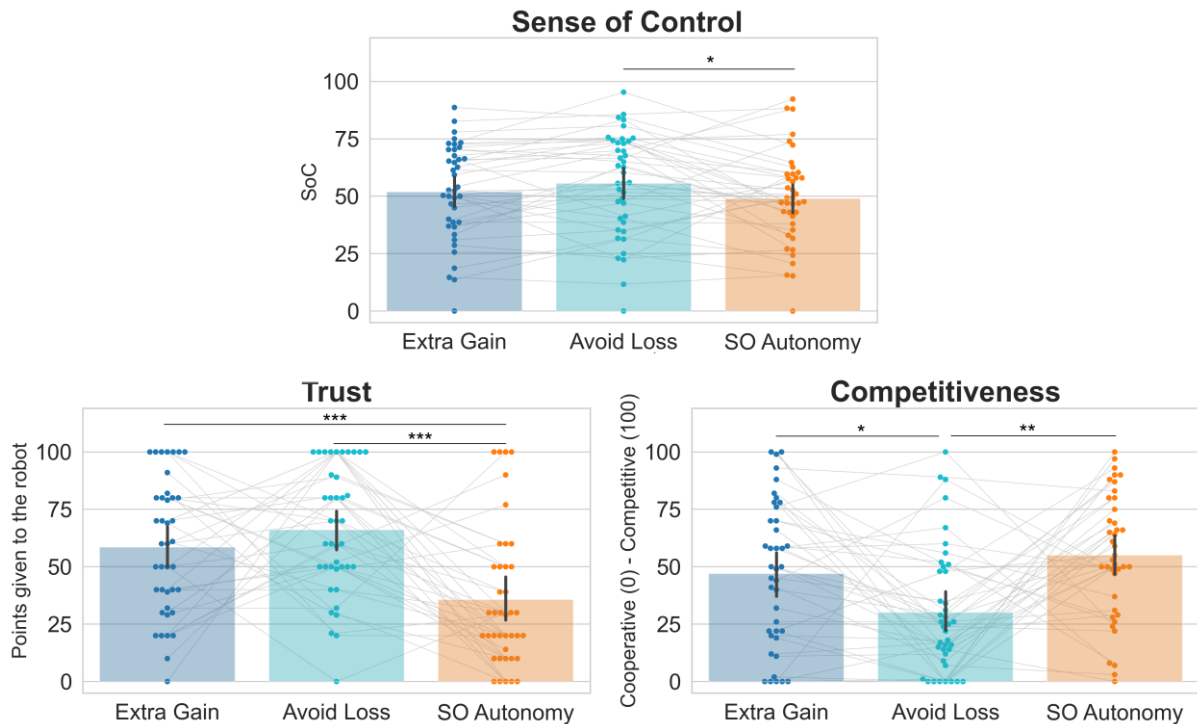
Finally, introducing two types of helpful autonomy meant that we had to use different payoff matrices than in the previous experiments. In contrast to previous experiments each block contained 12 trials (and not 10). In five trials the visible boxes displayed 0 and 5 points, in five trials they displayed -10 and -5 points, and in two trials they displayed 0 and 10 points.

In each trial the exact location of each option (left, left-center, right-center, right), including which box was hidden and visible, was random. Each robot behaved autonomously in 5 trials/block. The Autonomy Extra-Gain robot always chose 10 points when participants had to choose between 0 and 5. The Autonomy Avoid-Loss robot always chose 0 when participants had to select -10 or -5. The Same-Outcome Autonomy robot exhibited autonomy also 5 times by selecting a hidden option that was equal in payoff to the highest visible option. It did so twice for the [-10, -5] visible options, twice for the [0, 5], and once for the [0, 10].

## 4.2. Results

### 4.2.1. Preregistered analyses

**Sense of control:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on reports of sense of control ( $F(1.6,61.6)=3.4$ ,  $p=0.049$  Greenhouse-Geisser corrected, partial  $\eta^2=0.08$ ). Subsequent post hoc tests showed that sense of control was judged as significantly higher for Autonomy Avoid Loss than for Same-Outcome Autonomy ( $p=0.032$ ). The remaining differences were not statistically significant. Figure 7 illustrates the results.



**Figure 7.** Results of the preregistered analyses for Experiment 3. Statistically significant differences are indicated by stars. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < 0.001$ .

**Trust:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on the number of points given to a robot in the trust game ( $F(2,78)=13.6$ ,  $p<0.001$ , partial  $\eta^2=0.26$ ). Subsequent post hoc tests showed that trust in the Same-Outcome Autonomy was significantly lower than in the other two conditions (both  $p<0.001$ ). The difference between Autonomy Extra Gain and Autonomy Avoid Loss was not significant ( $p=0.22$ ).

**Competitiveness:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on judgments of perceived competitiveness ( $F(1.7,67.1)=6.5$ ,  $p=0.004$  Greenhouse-Geisser corrected, partial  $\eta^2=0.14$ ). Subsequent post hoc tests showed that Autonomy Avoid Loss was perceived as significantly less competitive than Same-Outcome Autonomy ( $p=0.002$ ) and Autonomy Extra Gain ( $p=0.039$ ). The difference between Same-Outcome Autonomy and Autonomy Extra Gain was not significant ( $p=0.26$ ).

In contrast to the previous experiments, none of the helpful autonomies had a clear bimodal distribution. However, Autonomy Avoid Loss had a distribution centered around lower numbers, while Autonomy Extra Gain had a flatter distribution with a peak close to the peak of the Autonomy Same.

## 4.3. Discussion

Experiment 3 found several differences between the investigated three types of robot autonomy, but also revealed differences when compared with the previous two Experiments. First, we replicated our previous findings that helpful autonomy leads to higher trust than non-helpful autonomy. At the same time, the difference between two types of helpful autonomy was not significant. This pattern of results did not translate to perceived competitiveness, where a robot showing autonomy to avoid loss was perceived as significantly less competitive than the remaining two robots (which did not significantly differ from each other). This supports our findings from the previous experiments that trust, as measured in the trust game, and perceived competitiveness are driven by different factors.

The most surprising finding from Experiment 3 was that reported sense of control was numerically higher for both types of helpful autonomy than for Same-Outcome Autonomy (although this difference was significant only for autonomy to avoid loss). This contrasts with the results of Experiments 1 and 2 where helpful autonomy led to a decrease in sense of control, relative to same-outcome autonomy. We propose two potential explanations for the reversal of this pattern. The first one is due to the reliability of robot behavior. In Experiment

3 robots demonstrating helpful autonomy were 100% reliable. It means that if a participant was at a risk of a loss (had to choose between -5 and -10 points) then the Avoid Loss robot would always intervene and choose an option leading to no loss (0 points). Similarly, if the highest available option to choose was 5 points then the Extra Gain robot would always find a hidden option with 10 points (the highest score one could achieve in a single trial). If participants learned these patterns of behavior, then they could use them to explain a huge part of the potential prediction error of the helpful robots just by seeing the available options. For example, seeing two negative options when controlling an Autonomy Avoid Loss robot might mean that whatever choice they make, the robot will always save them from loss and steer towards a hidden box with zero points. Such behavior would then lead to much smaller prediction error than in Same-Outcome Autonomy, where there was always uncertainty whether the robot would obey or disobey the participant's command.

The second explanation pertains to the factors that are outside of the behavior of that specific robot, but instead refer to the context of the task. It is not a single explanation, but a family of them. The two most important changes introduced in Experiment 3 were: dividing the helpful autonomy into two types, and removing the control conditions of full control and random behavior. If this explanation is correct, then the explicitly reported sense of control can be influenced by how it compares against other conditions and what type of control conditions are involved.

## **5. Experiment 4**

The goal of Experiment 4 was to dissociate between the two potential explanations of the surprising result in sense of control obtained in Experiment 3. Specifically, we decided to directly target the hypothesis that relative increase in the sense of control for helpful autonomy was caused by the fact that in Experiment 3 both types of helpful autonomy were 100% reliable, henceforth strongly reducing the magnitude of prediction error in these circumstances. To test this explanation, we conducted Experiment 4 which was a direct replication of Experiment 3, but this time helpful autonomy was only 60% reliable. It means that when, for example, a participant had to choose between two losing options (-5 and -10) and not making a choice would incur an even larger cost (-20), the robot helping to avoid loss would select the hidden option with 0 payoff only 60% of the time. The same applies to the robot that helped to obtain an extra gain. We predicted that if reliability was the driving factor of our results in Experiment 3 then in Experiment 4 both types of helpful autonomy should be

perceived as leading to lower sense of control than Same-Outcome Autonomy. Conversely, if we observe similar pattern of results to the one from Experiment 3 then we can conclude that it is caused by how sense of control in helpful autonomy is compared against SoC in other conditions.

## **5.1. Methods**

### **5.1.1. Preregistration**

The experiment has been preregistered. Preregistration form is available under the following link: [https://aspredicted.org/4R1\\_8DR](https://aspredicted.org/4R1_8DR)

### **5.1.2. Participants**

Calculations conducted with GPower 3.1.9.7 for the main effect of condition (F test for ANOVA with repeated measures,  $\alpha=0.05$ ,  $\beta=0.95$ , number of measurements=3) for a small-to-medium effect (effect size  $f=0.33$ ) indicated that the sample size should be at least 26 participants. To account for the fact that online studies can be characterized by noisy results we set our target sample size to 40 participants. The preregistered exclusion criteria were the same as in Experiment 1 and 2.

61 participants completed the experiment and provided full datasets. Three participants failed to respond in more than 10% of the trials and were excluded from the analysis. 18 participants incorrectly described the behavior of at least one robot, so new participants were tested in their place leading to the final sample of 40 participants. The results from 58 participants are presented in Supplementary Materials S1, but they do not substantially differ from the results described here.

Of the final 40 participants 14 were female and 26 male. The mean age was 32.7 years ( $SD=13$ , range: 18-77). The nationality of participants was: Poland (10), UK (7), Portugal (6), Italy, Hungary, South Africa (3), Mexico, Azerbaijan, Turkey, Bosnia and Herzegovina, Canada, Spain, Greece (1).

### **5.1.3. Design**

The design was identical to Experiment 3.

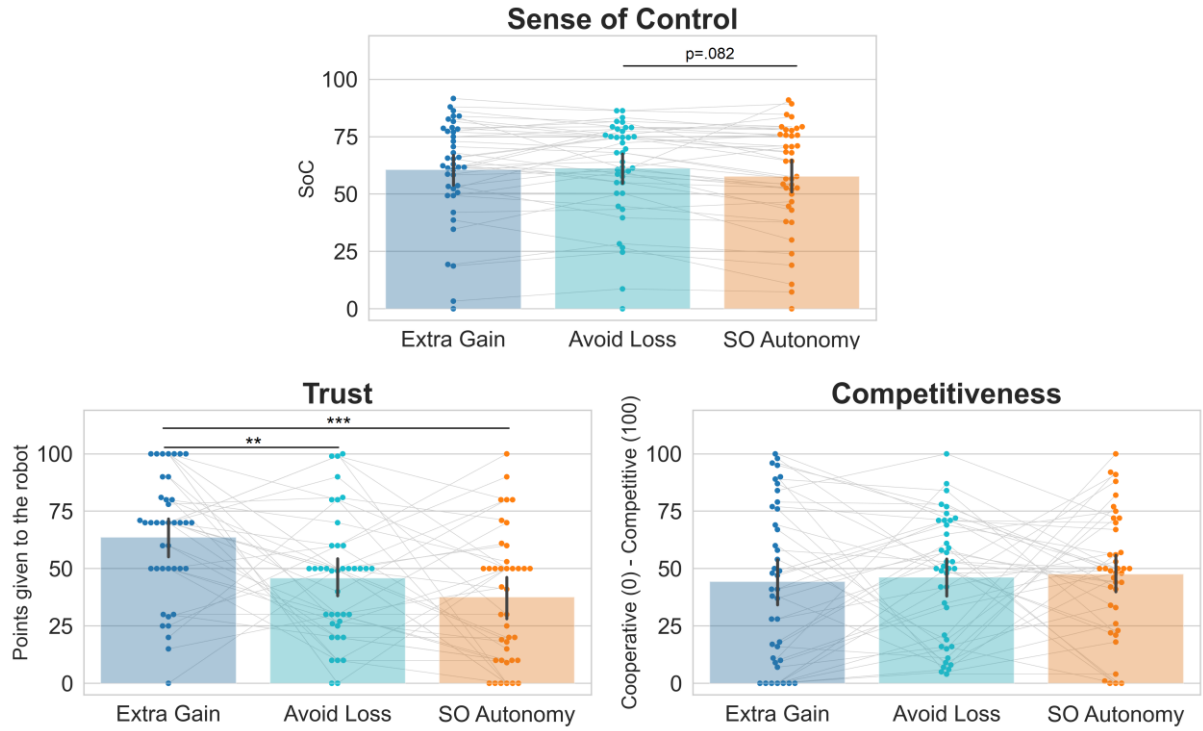
#### 5.1.4. Procedure

The procedure of the experiment was the same as in Experiment 3 with one difference: in each block each robot exhibited autonomy 3 times rather than 5 times (out of 12 trials). Specifically, for the Autonomy Extra Gain robot: in trials in which the visible options were 0 and 5 it disobeyed the participant's commands (and selected a hidden option with 10 points) in 3 out of 5 trials, and in 2 of such trials it followed the participant's command. For the Autonomy Avoid Loss the same logic applied regarding trials in which participants had to choose between -10 and -5 points. Out of 5 such trials: in 3 trials the robot selected 0 points and in 2 trials it followed the participant's instructions. Finally, the Same-Outcome Autonomy robot displayed autonomy in three trials per block, one per each type of trials with visible options [-10, -5], [0, 5], and [0, 10].

## 5.2. Results

### 5.2.1. Preregistered analyses

**Sense of control:** A one-way repeated-measures ANOVA revealed a marginally significant main effect of Robot Behavior on reports of sense of control ( $F(1.7, 67.0)=2.9$ ,  $p=0.070$  Greenhouse-Geisser corrected, partial  $\eta^2=0.07$ ). Subsequent post hoc tests showed that sense of control was judged as marginally higher for Autonomy Avoid Loss than for Same-Outcome Autonomy ( $p=0.082$ ). The remaining differences were not statistically significant ( $p>0.13$ ). Figure 8 illustrates the results.

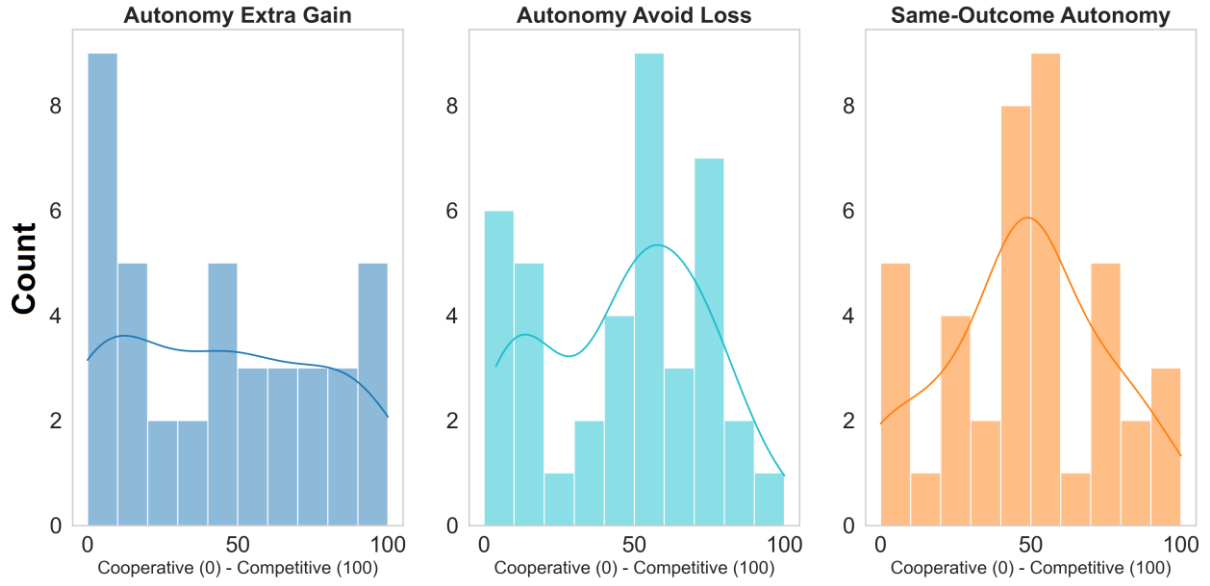


**Figure 8.** Results of the preregistered analyses for Experiment 4. Statistically significant differences are indicated by stars. \*\*  $p < .01$ , \*\*\*  $p < 0.001$ .

**Trust:** A one-way repeated-measures ANOVA revealed a significant main effect of Robot Behavior on the number of points given to a robot in the trust game ( $F(2,78)=10.9$ ,  $p < 0.001$ , partial  $\eta^2=0.22$ ). Subsequent post hoc tests showed that trust in the Autonomy Extra Gain was significantly higher than in the other two conditions ( $p=0.005$  with Autonomy Avoid Loss and  $p < 0.001$  with Same-Outcome Autonomy). The difference between Autonomy Avoid Loss and Same-Outcome Autonomy was not significant ( $p=0.15$ ).

**Competitiveness:** A one-way repeated-measures ANOVA did not reveal a significant main effect of Robot Behavior on judgments of perceived competitiveness ( $F(1.7,66.0)=0.12$ ,  $p=0.85$  Greenhouse-Geisser corrected, partial  $\eta^2 < 0.01$ ).

In contrast to Experiment 3, both types of helpful autonomies displayed a bimodal distribution of results, while Same-Outcome Autonomy once again revealed a unimodal distribution centered around the value of 50.



**Figure 9.** The distribution of ratings of competitiveness of each robot from Experiment 4.

### 5.3. Discussion

Experiment 4 was designed similarly to Experiment 3, but with both types of helpful autonomy being unreliable (the helpful robots helped 60% of the time rather than always). We expected that if 100% reliability (and hence full predictability of robot behavior) was the driving factor of higher sense of control for helpful than non-helpful autonomy in Experiment 3, then Experiment 4 should show a different pattern from that of Experiment 3, but a similar pattern to that of Experiments 1 and 2. This was not the case. The difference between Autonomy Same and the two types of helpful autonomy was not statistically significant and, in fact, numerically it even went in the opposite direction to this prediction (the difference between Same-Outcome Autonomy and autonomy to avoid loss was marginally significant). This pattern of results is a weakened pattern from Experiment 3 and makes it very unlikely that our first interpretation of the results of Experiment 3 was correct rather implausible. Instead, it provides a strong argument in favor of the explanation that this effect was due to the presence/absence of specific control conditions. There are previous reports showing that judgments of agency can be influenced by contextual factors (Baptista, Jacquet, Sidarus, Cohen, & Chambon, 2022). In our case, we speculate that the effect might be caused by the anchoring effect (Furnham & Boo, 2011; Tversky & Kahneman, 1974) of inclusion of a robot that is fully under one's control. In such case participants feel highest sense of control over a robot that always obeys their commands, but might feel more sense of control over the Same-Outcome Autonomy than the autonomy to choose better, because in same-outcome autonomy,



the robot's behavior is more similar to the behavior of the robot over which they have full control.

While robot reliability did not have a significant effect on sense of control, it did influence trust and perceived competitiveness. Interestingly, reliability influenced only the robot that helped to avoid loss. Specifically, if this robot was unreliable then trust towards it decreased to the level of a non-helpful robot, while trust towards a robot providing an extra gain remained at a significantly higher level. Second, while the Avoid-Loss robot in Experiment 3 was perceived as significantly more cooperative than the other two robots, this effect completely disappeared when it stopped being reliable, and started being perceived as equally competitive as the other two robots.

## **6. General Discussion**

The goal of our study was to investigate how different types of robot autonomy, with a special emphasis on helpful and same-outcome (neutral) autonomy, influence sense of control, trust, perceived competitiveness and sense of joint agency.

First, we discovered that under some circumstances people experience higher sense of control over helpful than neutral autonomous robots, and in other contexts this relationship reverses. This means that by manipulating contextual factors it is possible to increase or decrease reported sense of control over autonomous systems. In our case the most likely factor responsible for this pattern was presence or absence of other types of robots: one over which participant has full control and one that displays random behavior. We propose that presence of the first one might lead to an anchoring effect (Furnham & Boo, 2011; Tversky & Kahneman, 1974): participants could compare the behavior of autonomous robots with a robot that was fully under their control and interpret an autonomous robot that chooses a different option but with the same payoff (Same-Outcome Autonomy) as more similar to the fully controlled one due to the fact that it disobeys them only in one way (which option it chooses and not what number of points it gains), as compared to the helpful autonomous robot (which chooses a different option and acquires a different amount of points than those chosen by the participant). Conversely, when this comparison was not available, the effect of robot autonomy on sense of control went in the opposite direction: people felt more sense of control over the helpful robot than the Same-Outcome autonomous robot. In this case the lack of an anchor might have made the tendency to over-attribute positive outcomes to oneself stronger. There are also other alternative explanations pertaining to other differences between the first

two and the latter two experiments that cannot be fully ruled out (presence of one or two robots with helpful autonomy, presence or absence of a question about the sense of joint agency), but they all refer to factors that are external to the behavior of the robot. Overall, our finding suggests that it might be possible to increase a person's subjective experience of control by manipulating the context in which it is experienced. While illuminating the specifics of this effect in HRI requires further research, it might have important practical implications, as higher sense of control is typically associated with positive cognitive effects (higher attention and error monitoring etc.: Mulder et al., 2012; Navarro et al., 2016). At the same time, the type of helpful autonomy (avoiding loss or helping to obtain an extra gain) did not affect sense of control in our study.

Second, we found that participants experienced highest sense of joint agency (SoJA) with the robot that they fully controlled. This was a surprising finding, because we expected that it would be highest for a helpful autonomous robot. Instead, SoJA for helpful autonomous robots showed a clear bimodal distribution: participants either experienced very high joint agency (similar to the fully controlled robot in Experiment 1 and slightly lower in Experiment 2) or very low (lower than with the Same-Outcome Autonomy robots and only slightly higher than with a random robot). Sense of joint agency represents the feeling of doing an action **together** with another agent and hence requires both partners to contribute to the joint goal (Loehr, 2022; Shteynberg et al., 2023; Zapparoli et al., 2022, see also: Le Besnerais, Moore, Berberian, & Grynszpan, 2024). So high SoJA for fully controlled robots can be explained by the fact that participants represented these robots not as just tools but as agents being responsible for a complementary task: while participants were making the decisions, the robots executed them. As such they contributed to different aspects of the common task, and the fact that their responsibilities did not overlap eliminated any risk of misalignment. If this interpretation is correct, then the bimodal distribution for helpful autonomous robots might reflect two different ways of interpreting the situation in which a robot can intercept the decision-making process: either as an interruption of the otherwise smooth joint action (leading to low SoJA) or as a valuable contribution to it (leading to high SoJA). This interpretation is consistent with previous findings showing that any forms of disruptions or misalignment of intentions leads to the decrease of SoJA (Schwarz, Tonn, Büttner, Kunde, & Pfister, 2023; van der Wel, 2015; Wahn, Karlinsky, Schmitz, & König, 2018). However, it must be noted that in our study SoJA was strongly correlated with sense of control raising a potential concern that reports of SoJA could have been influenced by participants' reports of sense of control that were given just before.

Third, neither report of control nor of joint agency directly translated to trust towards robots, as measured with an economic trust game. We discovered that overall, people trusted most the robots displaying helpful autonomy. However, if the helpful autonomous robot led to the same payoff as the robot that was under full control then both of these robots were equally trusted (Experiment 2). Only when helpful autonomy allowed to reach even higher reward it was trusted the most (Experiment 1). This suggests that trust in the trust game might be especially related to whether we perceive the robot within the context of the task as efficient or competent in acquiring points. This interpretation finds support in a proposal by Malle & Ullman (2021) who distinguished two types of trust: performance and moral. Our data suggest that trust game results might target the performance trust, and it is unclear whether our results would replicate in a moral trust context, for example when telling a personal secret to a robot or letting a robot take care of one's beloved person. Second, we found that the type of helpful autonomy can affect trust towards a robot, but it depends on the reliability of helping behavior. Specifically, we found that participants trusted a robot that helped them to obtain an extra gain regardless of whether it did so reliably or not. However, for a robot that helped them to avoid a loss they showed an increased trust only if it was reliable in doing so. If its reliability was reduced (to 60% in our case) then trust towards it decreased to the level shown towards a non-helpful autonomous robot. This finding underscores the importance of taking humans' loss aversion (Abdellaoui, Bleichrodt, & Paraschiv, 2007; Sokol-Hessner & Rutledge, 2019; Tversky & Kahneman, 1992) together with its reliability (Desai et al., 2012; Wright, Chen, & Lakhmani, 2019; Zhang, Lee, Maeng, & Hahn, 2023) into consideration when developing autonomous helpers. The fact that this effect emerged also in our study in which loss of points did not have any tangible effect on participants (and in fact this robot obtained exactly the same cumulative reward as the one helping to obtain extra gain) illustrates the importance of reliability in the context of robots that help us to prevent or avoid negative events.

Finally, the findings related to trust and SoJA find further support in our results regarding the perceived competitiveness of a robot. On the one hand, robots displaying helpful autonomy showed a bimodal distribution of perceived competitiveness (similarly to SoJA). It suggests that such robots can be either perceived as very cooperative (non-competitive), or as highly competitive, and therefore potentially threatening. Moreover, robots that have the autonomy to help to avoid a loss are perceived as especially cooperative but only if they are reliable. Otherwise, they lose their edge – similarly to the pattern observed for trust.

## Conclusions

In our study we investigated how different forms of robot autonomy influence sense of control and trust towards the robots. First, we found that people tend to trust an autonomous robot if they see that it helps them to achieve their goals. However, it is also important to take into account the nature of these goals. If it is gaining an additional unexpected reward then people appreciate such help even if it is unreliable. On the other hand, if the task of the robot is to help avoid a potential loss then the issue of reliability becomes important – if it is unreliable then it significantly decreased trust in such robot. Second, sense of control over the robot appears to be largely unrelated to trust towards it, although our study discovered that in the case of autonomous robots it can be influenced by factors going beyond the single robot's behavior, but also by how it compares to other available robots and their behaviors. Finally, our results suggest that helpful autonomous robots tend to be perceived in one of two ways: either as highly cooperative (what is also reflected by high sense of joint agency when doing a task with them) or as a potentially competitive and perhaps even threatening.

## Supplementary Materials

S1. Full data analysis report for confirmatory analyses

S2. Full data analysis report for confirmatory analyses with the excluded participants

S3. Results of exploratory correlational analyses

S4. Experimental materials and data analysis scripts are available under the following link:

<https://osf.io/3ng89>

## References

- Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007). Loss aversion under prospect theory: A parameter-free measurement. *Management science*, 53(10), 1659-1674.
- Banerjee, S., Galizzi, M. M., & Hortala-Vallve, R. (2021). Trusting the trust game: An external validity analysis with a UK representative sample. *Games*, 12(3), 66.
- Baptista, A., Jacquet, P. O., Sidarus, N., Cohen, D., & Chambon, V. (2022). Susceptibility of agency judgments to social influence. *Cognition*, 226, 105173.
- Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of human-robot interaction*, 3(2), 74.
- Berberian, B., Sarrazin, J.-C., Le Blaye, P., & Haggard, P. (2012). Automation technology and sense of control: a window on human agency. *PLoS One*, 7(3), e34075.
- Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nat Neurosci*, 1(7), 635-640. doi:10.1038/2870

- Camerer, C. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. *Available at SSRN 1977749*.
- Cornelio, P., Haggard, P., Hornbaek, K., Georgiou, O., Bergström, J., Subramanian, S., & Obrist, M. (2022). The sense of agency in emerging technologies for human–computer integration: A review. *Frontiers in neuroscience*, 16, 949138.
- De Dreu, C. K., Gross, J., Arciniegas, A., Hoenig, L. C., Rojek-Giffin, M., & Scheepers, D. T. (2024). On being unpredictable and winning. *Journal of Personality and Social Psychology*, 126(3), 369.
- De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1-12.
- Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., . . . Yanco, H. (2012). *Effects of changing reliability on trust of robot systems*. Paper presented at the Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.
- Dewey, J. A., Seiffert, A. E., & Carr, T. H. (2010). Taking credit for success: The phenomenology of control in a goal-directed task. *Consciousness and cognition*, 19(1), 48-62.
- Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3), 462-492.
- Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1404), 1771-1788.
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The journal of socio-economics*, 40(1), 35-42.
- Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: a systematic lab-field study. *Management science*, 65(3), 976-1002.
- Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn Sci*, 4(1), 14-21. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/10637618>
- Grynszpan, O., Sahaï, A., Hamidi, N., Pacherie, E., Berberian, B., Roche, L., & Saint-Bauzel, L. (2019). The sense of agency in human-human vs human-robot joint action. *Consciousness and cognition*, 75, 102820.
- Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4), 196.
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5), 517-527.
- Jammes, Y., Behr, M., Llari, M., Bonicel, S., Weber, J. P., & Berdah, S. (2017). Emergency braking is affected by the use of cruise control. *Traffic injury prevention*, 18(6), 636-641.
- Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99-127): World Scientific.
- Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12, 604977.
- Kok, B. C., & Soh, H. (2020). Trust in robots: Challenges and opportunities. *Current Robotics Reports*, 1, 297-309.
- Le Besnerais, A., Moore, J. W., Berberian, B., & Grynszpan, O. (2024). Sense of agency in joint action: A critical review of we-agency. *Frontiers in psychology*, 15, 1331084.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.
- Legaspi, R., & Toyoizumi, T. (2019). A Bayesian psychophysics model of sense of agency. *Nature communications*, 10(1), 1-11.
- Levitt, S. D., & List, J. A. (2007). On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d'économique*, 40(2), 347-370.
- Limerick, H., Coyle, D., & Moore, J. W. (2014). The experience of agency in human-computer interactions: a review. *Frontiers in human neuroscience*, 8, 643.

- Loehr, J. D. (2022). The sense of agency in joint action: An integrative review. *Psychonomic bulletin & review*, 1-29.
- Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction* (pp. 3-25): Elsevier.
- Moore, J. W. (2016). What is the sense of agency and why does it matter? *Frontiers in psychology*, 7, 1272.
- Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., . . . Legg, S. (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462*.
- Mulder, M., Abbink, D. A., & Boer, E. R. (2012). Sharing control with haptics: Seamless driver support from manual to automatic control. *Human factors*, 54(5), 786-798.
- Navare, U., Ciardo, F., Kompatsiari, K., De Tommaso, D., & Wykowska, A. (2023). Performing actions with robots: attribution of intentionality affects the joint sense of agency. *International Journal of Social Robotics*, 12(6), 1203-1211.
- Navarro, J., François, M., & Mars, F. (2016). Obstacle avoidance under automated steering: Impact on driving and gaze behaviours. *Transportation research part F: traffic psychology and behaviour*, 43, 315-324.
- Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, 13(1), 1-30.
- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179-217.
- Pagliari, M., Chambon, V., & Berberian, B. (2022). What is new with Artificial Intelligence? Human-agent interactions through the lens of social agency. *Frontiers in psychology*, 13, 954444.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(3), 286-297.
- Pesquita, A., Whitwell, R. L., & Enns, J. T. (2018). Predictive joint-action model: A hierarchical predictive approach to human cooperation. *Psychonomic bulletin & review*, 25, 1751-1769.
- Schwarz, K. A., Tonn, S., Büttner, J., Kunde, W., & Pfister, R. (2023). Sense of agency in social hierarchies. *Journal of Experimental Psychology: General*.
- Selvaggio, M., Cognetti, M., Nikolaidis, S., Ivaldi, S., & Siciliano, B. (2021). Autonomy in physical human-robot interaction: A brief survey. *IEEE Robotics and Automation Letters*, 6(4), 7989-7996.
- Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: measurement and application to system design. *Frontiers in psychology*, 10, 1117.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Retrieved from
- Shteynberg, G., Hirsh, J. B., Wolf, W., Bargh, J. A., Boothby, E. B., Colman, A. M., . . . Rossignac-Milon, M. (2023). Theory of collective mind. *Trends in Cognitive Sciences*.
- Sokol-Hessner, P., & Rutledge, R. B. (2019). The psychological and neural basis of loss aversion. *Current directions in psychological science*, 28(1), 20-27.
- Sottysik, M., Gawłowska, M., Sniezynski, B., & Gunia, A. (2024). *Artificial Intelligence, Management, and Trust*: Routledge.
- Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious Cogn*, 17(1), 219-239. doi:10.1016/j.concog.2007.03.010
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124-1131.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of risk and uncertainty*, 5, 297-323.
- van der Wel, R. P. (2015). Me and we: Metacognition and performance evaluation of joint actions. *Cognition*, 140, 49-59.
- Wahn, B., Karlinsky, A., Schmitz, L., & König, P. (2018). Let's move it together: A review of group benefits in joint object control. *Frontiers in psychology*, 9, 918.

- Wen, W., Brann, E., Di Costa, S., & Haggard, P. (2018). Enhanced perceptual processing of self-generated motion: Evidence from steady-state visual evoked potentials. *Neuroimage*, 175, 438-448.
- Wen, W., & Imamizu, H. (2022). The sense of agency in perception, behaviour and human–machine interactions. *Nature Reviews Psychology*, 1(4), 211-222.
- Wen, W., Kuroki, Y., & Asama, H. (2019). The sense of agency in driving automation. *Frontiers in psychology*, 10, 2691.
- Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural networks*, 11(7-8), 1317-1329.
- Woźniak, M., & Knoblich, G. (2022). Communication and action predictability: two complementary strategies for successful cooperation. *PsyArXiv*. doi:10.31234/osf.io/sx24v
- Wright, J. L., Chen, J. Y., & Lakhmani, S. G. (2019). Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on human-machine systems*, 50(3), 254-263.
- Wykowska, A. (2021). Robots as mirrors of the human mind. *Current directions in psychological science*, 30(1), 34-40.
- Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693), 20150375.
- Yoshie, M., & Haggard, P. (2013). Negative emotional outcomes attenuate sense of agency over voluntary actions. *Current Biology*, 23(20), 2028-2032.
- Zapparoli, L., Paulesu, E., Mariano, M., Ravani, A., & Sacheli, L. M. (2022). The sense of agency in joint actions: a theory-driven meta-analysis. *Cortex*.
- Zhang, X., Lee, S. K., Maeng, H., & Hahn, S. (2023). Effects of Failure Types on Trust Repairs in Human–Robot Interactions. *International Journal of Social Robotics*, 15(9), 1619-1635.