Leveraging human agency to improve confidence and acceptability in human-machine interactions

Quentin Vantrepotte, Bruno Berberian, Marine Pagliari, Valérian Chambon

Leveraging human agency to improve confidence and acceptability in human-machine interactions

Quentin Vantrepotte[1,2*], Bruno Berberian[1*], Marine Pagliari[1,2], Valérian Chambon[2*]

[1] Institut Jean Nicod, Département d'Études Cognitives, École Normale Supérieure, CNRS, PSL University, Paris, France.

[2] Information Processing and Systems, ONERA, Salon de Provence, Base Aérienne 701, France

*Corresponding authors: Quentin Vantrepotte (quentin.vantrepotte@onera.fr); Bruno Berberian (bruno.berberian@onera.fr); Valerian Chambon (valerian.chambon@gmail.com).

Abstract

Repeated interactions with automated systems are known to affect how agents experience their own actions and choices. The present study explores the possibility of partially restoring sense of agency in operators interacting with automated systems by providing additional information about the system's decision, i.e. its confidence. To do so, we implemented an obstacle avoidance task with different levels of automation and explicability. Levels of automation were varied by implementing conditions in which the participant was free or not free to choose which direction to take, whereas levels of explicability were varied by providing or not providing the participant with the system's confidence in the direction to take. We first assessed how automation and explicability interacted with participants' sense of agency, and then tested whether increased self-agency was systematically associated with greater confidence in the decision and improved system acceptability. The results showed an overall positive effect of system assistance. Providing additional information about the system's decision (explicability effect) and reducing the cognitive load associated with the decision itself (automation effect) was associated with stronger sense of agency, greater confidence in the decision, and better performance. In addition to the positive effects of system assistance, acceptability scores revealed that participants perceived "explicable" systems more favorably. These results highlight the potential value of studying self-agency in human-machine interaction as a guideline for making automation technologies more acceptable and, ultimately, improving the usefulness of these technologies.

Keywords*:* Sense of Agency; Confidence; Temporal binding; Human-machine interaction; Explicability; Acceptability.

1 Introduction

Human agents are used to interacting with sophisticated computer systems designed to help them in their activities and a significant number of our daily actions are technologically mediated. This is the case of the airplane pilot who controls their aircraft through increasingly sophisticated pilot assistance tools, but also that of the doctor assisted in their diagnoses by artificial intelligence algorithms. This paramount importance of technological assistance is further demonstrated by the growing role that virtual assistants – be it our phones, computers or specific planning devices – play in our daily lives.

Repeated interactions with automated systems can be expected to affect the way individuals experience their own actions and choices. Such experience is often referred to as "sense of agency" (SoA) and describes the subjective feeling associated with controlling one's own actions and, through these actions, events in the outside world (Haggard and Tsakiris, 2009). SoA plays a key role in guiding attributions of responsibility (Bigenwald and Chambon, 2019) and serves as a key motivational force for human behaviour (di Costa et al., 2018). A detrimental effect of human-machine interactions on SoA has been demonstrated in a number of experimental studies, one of the most typical of which implements an aircraft supervision task (e.g. Berberian et al., 2012). In this task, the participant is responsible for supervising the movement of an aircraft that may encounter unpredictable obstacles. When a conflict occurs due to the presence of another aircraft, the participant has to decide and implement the appropriate control to avoid the obstacle using a button-based interface. Following an established classification (Sheridan & Verplank, 1978), the levels of automation of the task were manipulated from the user having complete control (no automation) to the computer performing the entire task with the participant simply observing (full automation). The results showed a decrease in the participant's sense of agency concomitant with the increase in automation, and suggest that increasing the level of automation tends to distract operators from the results of the action and alter the emergence of a sense of control.

What makes automation particularly detrimental to the operator's sense of agency is not yet fully understood, but there is a relative consensus that the lack of transparency on how the system makes its decisions, or simply operates, is a key factor (Christoffersen & Woods, 2002; Klien et al., 2004). In most human-machine interactions, most of these decision processes are unknown, inaccessible, or even not explainable at all (Norman, 1990). Such opacity makes it difficult for the operator to link system intention to actual state and to predict the sequence of events that will occur. Predictive mechanisms are known to play a critical role in the development of individuals' sense of agency, by allowing the attribution of observed sensory events to prior intentions (Haggard & Eitam,

2015; Chambon et al., 2013; Chambon, Sidarus, Haggard, 2014). By affecting the predictability of actions, the inherent opacity of technological systems is therefore likely to alter perceived agency in human operators.

In a previous study, we directly investigated how system predictability impacts the development of agency experience during human-machine interaction (Le Goff et al, 2018). Particularly, we explored the benefit of prime messages regarding system intention while supervising an automated system. We tested whether providing information about what to do next mitigated the deleterious effect of reduced freedom of action on agency and, in doing so, increased the user's level of acceptability, along with increased control and performance. Our results suggest that displaying the system's intentions prior to an action is a good candidate for maximizing the experience of agency in supervisory task, and for increasing system acceptability as well. These preliminary results open interesting avenue as to how to modulate the emergence of the experience of agency during human-machine interaction.

The present study aims to go further in exploring the information required for making technological systems more intelligible, and to test whether improving intelligibility concomitantly increases the level of agency experienced by human operators. In two distinct experiments, we explored the role of communicating specific metacognitive information on improving the SoA of participants interacting with distinct automated systems. Specifically, both experiments implemented an avoidance task with different levels of automation and explicability.

Levels of automation were varied by implementing conditions in which the participant was free to choose which direction to take (free choice trials) or not (forced choice trials). Levels of explicability were varied by providing or not providing the participant with the system's confidence in the direction to take. Confidence can be seen as a measure of the uncertainty (or certainty) associated with one's choice or action (Fleming et al., 2014). Communicating confidence was intended to improve explicability of the system's decision, by increasing its transparency, that is, by providing the participant with additional information, such as the level of confidence associated with that decision (Tintarev and Masthoff, 2015). Indeed, the level of uncertainty (or confidence) associated with a decision is a key explanatory factor for why a decision is made or not, and whether or not that decision will be updated or revised in the future (Balsdon et al., 2020). The beneficial role of confidence on decision making has already been demonstrated in group settings, where sharing metacognitive representation increases joint performance (Bahrami et al. 2010; Fusaroli et al., 2012; La Bars et al., 2020) and enhances team coordination (Poizat et al., 2009; Lausic et al., 2009; Le Bars et al., 2021). Communicating confidence also makes performance more fluid and prospectively improves SoA (Sidarus et al., 2018; Chambon, Filevitch, Haggard, 2014), especially when

sensorimotor information is not available (Pacherie, 2013) such as when interacting with an automated system. Finally, there is indirect evidence that improving the operator's SoA during interaction with automated systems concomitantly improves acceptability of the system's decision itself (Le Goff et al., 2018). In addition to exploring the relationships between explicability and SoA, we also tested whether an increase in the participant's SoA could be consistently associated with greater confidence in their decision and greater acceptability of the system.

In both experiments, the participant's choice and three additional measures were collected: (i) Temporal Binding (TB), a widely-reported temporal compression between a voluntary action and its consequence (hence originally referred to as "intentional binding"), as an *implicit* proxy of the participant's sense of agency (Haggard, Clark, & Kalogeras, 2002; Caspar et al., 2017; Vogel et al., 2021; Ebert and Wegner, 2010); (ii) a measure of the participant's confidence in either their decision (free trials) or the system's decision (forced trials); and finally, (iii) the perceived acceptability of the system by the participant (Van der Laan et al., 1997).

In Experiment 1, we had three key predictions: (1) Participants would experience lower levels of agency when forced to follow the system's decision, compared to freely choosing (automation effect) (Berberian et al., 2012; Barlas et al., 2018; Caspar et al., 2018); (2) communicating to participants the system's confidence in the best decision would restore or even improve participants' SoA (explicability effect) (Sidarus et al., 2017); and finally (3) an increase in the participant's SoA would be associated with an increase in system acceptability (Le Goff et al., 2018). In Experiment 2, we further explored the relationships between our control measure, our decision confidence measure, and task demands. We leveraged a procedure developed in a previous study (Potts and Carlson, 2019) to clarify the contribution of task difficulty to the relationship between automation, explicability and sense of agency, using a modified version of the avoidance task from Experiment 1.

2 Experiment 1

2.1 Methods

2.1.1 Participants

Forty-four participants were recruited to participate in Experiment 1 (31 females, mean age = 33.2, SD = 8.4). In the absence of existing data with regard to our research goal, sample size was determined a priori on the basis of previous studies on SoA using temporal binding measures in a similar experimental design (free vs. forced-choice trials, Caspar et al, 2017). With this in mind, we targeted a sample size of 44 participants, similar to that of Caspar and colleagues (2017), with a potential dropout/exclusion rate of about 10% (in practice, 7 subjects were excluded on the basis of our exclusion criteria). The study was approved by the local ethics committee (reference: IRB n. 21-810). All participants gave written informed consent before inclusion in the study, which was carried out in accordance with the declaration of Helsinki (1964, revised 2013). The inclusion criteria were being older than 18 years, reporting no history of neurological or psychiatric disorders, no auditory disorders, and a normal or corrected-to-normal vision. Participants were all naive to the purpose of the study.

The studies were conducted on the experimental platform (PRISME) of the Institut du Cerveau et de la Moelle épinière (ICM) in Paris. The participants were tested in groups (maximum 12 participants) in a large testing room designed for this purpose. Each participant was brought individually in front of a computer, isolated by partitions, and was equipped with noise-cancelling headphones. The study was divided into two sessions, on two different days to reduce the effect of fatigue. Each session lasted 2 hours, and participants were paid 40 euros after completing both sessions. The procedure was the same for both sessions 1 and 2 (training and calibration phases, followed by the experimental task).

2.1.2 Material and stimuli

The participants were seated in front of a screen at approximatively 60cm (HP 23i; resolution: 1980x1080, 60 Hz). A keyboard and a noise cancelling headphone were used to perform the experiment.

The protocol used is the combination of an avoidance task, derived from the experimental paradigm Random Dot Kinematogram (RDK), followed by a temporal estimation task between an action and its effect (Temporal Binding measure). Matlab R2016b (MathWorks Inc.) and the Psychophysics Toolbox (Brainard 1997; Pelli 1997; Kleiner et al., 2007) were used to write and run

the task. RDKs were implemented using a sequence of white random dots that appeared briefly (800ms) within a circular aperture of 8° diameter. The white dots (1000 in total) were 4-by-4 pixels (0.107° square) and they were presented on a black background. The orientation of the target cloud was 15° along the vertical axis (195°, i.e. 180°+15°: leftward orientation; 165°, i.e. 180°-15°: rightward orientation). The coherence of the dots varied randomly: as the coherence decreased, a proportion of "non-coherent" dots, moving randomly, appeared. Coherence varied between 0 (all dots moved randomly) and 1 (all dots moved coherently with the orientation of the "target cloud"). The starting position of each dot in the cloud was randomly generated at the beginning of each trial. The type of trials and difficulty levels, as well as the performance of the system (error vs. correct) and the intervals presented, were fully balanced within each block. For each participant, guided and unguided blocks were presented in an alternating order, and this order (guided first, unguided next, or vice versa) was counterbalanced across participants.

2.2 General procedure

2.2.1 Training and calibration phases

Each session was split in two experimental blocks. The first block consisted of two training sessions and one calibration phase. The training sessions were designed to familiarize the participant with the avoidance task and the temporal estimation task, whereas the calibration phase was used to estimate a psychometric curve for each participant based on RDKs.

First training. The first training session consisted of the avoidance task alone. In this task, participants were presented with a screen consisting of a 2D aircraft silhouette placed in front of a moving dot cloud (see **Figure 1**). The cloud consisted of a majority of dots moving in one direction (target dots) and a few dots (non-coherent dots) moving in the other direction. The participant was instructed to detect the general orientation of the target cloud (left or right), which required ignoring the non-coherent dots. Once the orientation of the cloud detected, the participant was instructed to "avoid" it by making a left or right button press (e.g., left button press if the cloud appeared to be moving to the right). Note that the RDK was presented as an instantiation of a real-life flying situation ("You are flying a plane and there is a cloud of obstacles in front of you that you must avoid"). Thus, it was made clear that the kinematogram was a stimulus generated by a computer program, separate from the decision support system itself. At the end of each trial, feedback was given to the participant on whether or not they had successfully avoided the dot cloud. Twenty-four trials were presented, with different levels of coherence for the dot cloud. At the end of the training phase, participants who did not reach 75% correct responses had to repeat the 24 trials.

Calibration. The first training session was followed by a calibration phase, which was designed to estimate the participant's performance to the avoidance task for different levels of coherence by fitting a psychometric curve to each individual's data. The participant's task was the same as in the previous phase, except that there was no feedback on performance. To fit the psychometric curve, a double staircase method was used (Leek, 2001). A first scale started at maximum coherence and another scale started at the lowest level of coherence. The two scales were composed of 60 trials each; they alternated randomly and were independent of each other. The estimation was adapted to the participant's performance: when a participant achieved 3 successful trials of a series in a row (or 2 in half a series), the level of coherence decreased. As soon as a participant made a mistake, the level of coherence increased. The increase (or the decrease) was computed via a predefined step (equal to 0.8). At the end of the two series, a psychometric curve was computed. An adjustment with a logistical function was performed to improve the accuracy of the curve. Three levels of coherence, corresponding to three levels of difficulty on the avoidance task, were extracted from each participant's psychometric curve (easy, medium, and high levels corresponding to 100, 75, and 55% correct responses, respectively).

Second training. After the calibration phase, the participant conducted a second training session. This session consisted of the avoidance task, followed by a temporal estimation task. Specifically, the button press made in the avoidance task (left or right) was followed by a neutral sound for 300ms. Participants were informed that the time interval between their response and the sound could vary between 1 and 1500ms. Participants were presented with 15 pseudo-random latencies (equally distributed between 1ms and 1500ms) and gave their response by typing the estimated interval on the keyboard. The participant then received feedback on the actual delay and the next trial started. This interval judgment task was used to characterize a phenomenon known as "Temporal Binding" (TB), which refers to the perceived *compression* of the temporal interval between a voluntary action and its external consequence (Haggard, Clark, & Kalogeras, 2002). This compression is often seen an implicit marker of the sense of agency: a shorter perceived interval would indicate a higher sense of agency over the subsequent outcome (for a review, see Moore & Obhi, 2012). We computed TB as the difference between the perceived time and the actual delay between the button press and the sound, as routinely done in studies using the TB measure.

Testing phase

The testing phase consisted of the avoidance task followed by the temporal estimation task. A total of 12 blocks of 36 trials performed by each participant (432 trials in total). Three levels of difficulty,

corresponding to the 3 levels estimated from the calibration phase, were used during the avoidance task.

Avoidance task: automation factor. Two types of trials were presented, in which the level of automation of the response was manipulated. In half of the trials, participants were free to select the target arrow (left or right) that they believed was associated with the orientation of the cloud. In the other half, the computer randomly preselected an arrow, forcing the participant to match the computer's choice. Note that the choice imposed by the system may or may not be the same as the one the participant would have made spontaneously. Free- and forced-choice trials were randomly interleaved within each block. At the beginning of each trial, a brief sentence signaled the nature of the forthcoming condition (free trials: "You choose"; forced trials: "The system chooses") so as to avoid "anticipatory" free responses from the participant (see Chambon et al., 2020; Sidarus et al., 2019, for a similar procedure). Note however that the system was not fully autonomous as in other studies operating similar decision support systems (e.g., Le Goff et al., 2018), so any comparison with these studies should be viewed with caution. The system's choice and its confidence were always aligned – i.e. the system always chose the direction in which it was most confident.

Avoidance task: explicability factor. In addition to manipulating the level of automation over the decision, the level of *explicability* was also manipulated during the task. To do so, two decision support systems were implemented. In half of the trials, one system (system A) provided its *confidence* about its decision, while in the other half no confidence was provided (system B). Trials with confidence were labelled "guided trials" whereas trials without confidence were labelled "unguided trials" (see **Figure 2**). The confidence returned by system A could range from 0 (the system was not confident in its decision) to 100 (the system was absolutely certain of its decision). Importantly, the confidence provided by the system A was calibrated according to the difficulty of the current trial (metacognitive calibration). Thus, the system returned high confidence (95% confident on average ± 2 s.d. from the mean) for the chosen direction in easy trials, while returning lower confidence for medium (80% confident ± 2 s.d.) and difficult trials (65% certainty ± 2 s.d.). Note that system A returned its confidence even in trials where the subject could choose the direction on their own (free-choice trials, see **Figure 2**). Participants interacted with only one system within each experimental block (A or B).

The response reliability of each system was maintained at a high level, i.e. the system's performance was always better than that of the participants, except in easy trials where the performance of the system was identical to that of the subjects (100% correct answers). Performance of the systems was calibrated using the psychometric curve fitted to each participant's data in the previous phase. Thus, in easy, medium and difficult trials, the systems had 100%, 95% and 75% correct responses, respectively (for 100%, 75% and 55% correct responses in participants). The systems were slightly better than the participant so that they would be perceived as a real help, and to reduce any potential egocentric discounting bias (whereby participants automatically minimize external/social information, see Morin et al., 2021).

Time estimation task. As in the second training phase, the avoidance task was immediately followed by a time estimation task. Again, the task consisted of estimating the perceived time interval between the response to the orientation of the cloud and a subsequent sound. Participants were informed that the time interval between their response and the sound could vary between 1 and 1500ms, although only three intervals were used (250ms, 750ms and 1250ms).

Decision confidence. Following the time estimation task, the participant was asked to indicate confidence in their performance on the avoidance task using a Likert scale ranging from 1 (low confidence: the participant thinks they did not avoid the cloud) to 8 (high confidence: the participant thinks they avoided the cloud). Depending on the type of trial, the participant was asked to judge

either their own performance (free trial) or the performance of the system (forced trial) (see **Figure 2**). The participant indicated their response using the numeric keys on the keypad. Following the confidence rating, a feedback (green tick or red cross for correct or incorrect responses) was given to the participant on their performance on the avoidance task (**Figure 1**).

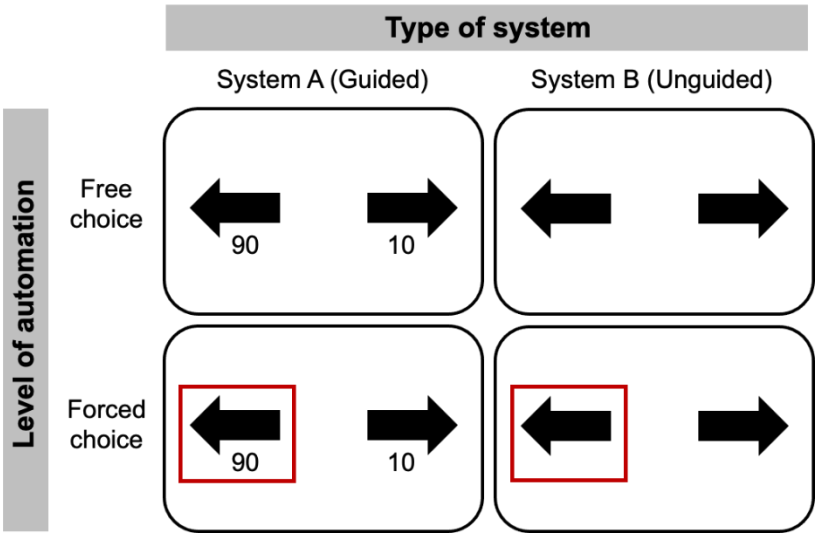

Figure 2. Illustration of the different types of response conditions. The free choice condition corresponds to trials in which the subject can choose the direction by herself. The forced choice condition corresponds to trials in which the subject must follow the choice of the system (indicated by a red square). One of the two systems (system A) guides the subject by returning its relative confidence (between 0 and 100) in each of the two possible answers. The second system returns nothing (no guidance).

Acceptability measure. Finally, at the end of each experimental block, participants were instructed to assess the "acceptability" felt toward the system (A or B) with which they had interacted. Two dimensions were measured to characterize such acceptability: Usefulness and Satisfaction. Usefulness was presented as the degree to which participants would expect to use the system if they had to perform a similar task, while Satisfaction was presented as the participant's satisfaction with the system (Van der Laan et al., 1997). Participants completed their evaluation by moving a cursor along two successive scales assessing either Usefulness or Satisfaction (not at all/totally).

2.3 Data analysis

Data were analysed using the R software (R Core Team, 2017) and the ezANOVA package (Lawrence, 2016). For the TB measure, the raw data were first filtered according to the interval reported by the participant for each interval category shown (250ms, 750ms and 1250ms). In each

participant, an average interval was calculated for each category and intervals that were two standard deviations above or below this average interval were considered outliers. Based on a pilot test with 16 participants, participants with more than 7.5% outliers were excluded from further analysis (2 participants). In order to further test the reliability of the reported intervals, we also calculated the correlation between perceived and actual intervals for each interval category and excluded 7 participants with non-significant coefficients (R's < 0.2) for the experiment 1.

Performance to the avoidance task (% of correct responses and response times) was analysed using a 2 × 3 repeated-measures ANOVA with explicability (guided vs. unguided) and task difficulty (easy vs. medium vs. high) as within-participants factors, whereas *TB* and *confidence* scores were analysed using 2 × 2 × 3 repeated-measures ANOVAs with automation (free vs. forced-choice), explicability (guided vs. unguided) and task difficulty (easy vs. medium vs. high) as within-participants factors. Note that percentages of correct responses were only analysed in the free-choice trials, where participants made an intentional choice. Main and interaction effects from all ANOVAs were further analyzed using Bonferroni-corrected post hoc pairwise comparisons. The datasets used for the analyses are: "*All_data_Exp1.csv*", "*CorrectRates_Experiment1.csv*" and "Acceptability_Exp1.csv" (Vantrepotte et al., 2021).

3 Results

3.1 Performance on the avoidance task - % Correct Responses

The ANOVA revealed significant main effects of the explicability (mean correct rates, guided = 87.88, SD = 11.59; mean correct rates, unguided = 81.71, SD = 15.42, $F(1,36) = 43.01$; $p < 0.001$; $\eta_p^2 = 0.544$) and the difficulty factors ($F(2,72) = 322,89$; $p < 0.001$; $\eta_p^2 = 0.900$). Post hoc comparisons further showed that participants gave more correct answers as the difficulty of the task decreased (all p's < 0.001). A significant explicability-by-difficulty interaction was also found ($F(2,72) = 9.61$; $p < 0.001$; $\eta_p^2 = 0.211$). Post hoc comparisons showed that the difference between guided and unguided trials increased with the difficulty (easy trials: $p = 0.209$; medium trials: $p < 0.001$; difficult trials: $p < 0.001$). These results suggest that our staircase procedure was reliable for assessing thresholds of difficulty in participants. They also show that the systems were, on average, much better than participants in deciding the right direction to take – in other words, the systems were generally very helpful to the participant during the experiment.

3.2 Temporal Binding

We found significant main effects of explicability (guided: –118.64 ± 340 vs. unguided: –86.71 ± 333, F(1,36) = 23.78; $p < 0.001$; $\eta_p^2 = 0.398$; left chart on the **Figure 3**), automation (F(1,36) = 4.90; $p = 0.03$; $\eta_p^2 = 0.120$) and difficulty (F(2,72) = 8.22; $p < 0.001$; $\eta_p^2 = 0.186$) on the TB measure. Bonferroni-corrected post hoc comparisons showed that the most difficult trials (high difficulty: –115.53 ± 339 vs. medium: –98.94 ± 336 vs. easy: –93.31 ± 336, all *p*'s < 0.001) were associated with a stronger binding effect, i.e. shorter action-outcome intervals.

A significant automation-by-difficulty interaction was also found (F(2,72) = 6.40; $p = 0.003$; $\eta_p^2 = 0.151$). Interestingly, free and forced choice trials differed only for the most difficult trials, with forced choices being associated with stronger binding effect than free choices (**Figure 3,** right panel). No other interaction effect was significant (automation-by-explicability: F(2,72) = 2.32; $p = 0.136$; explicability-by-difficulty interaction: F(2,72) = 2.25; $p = 0.113$). Bonferroni-corrected post hoc comparisons showed that the binding effect increased with difficulty in forced-choice trials, as opposed to free-choice trials (all $p > 0.05$), suggesting that task difficulty interacts with sense of agency for high levels of automation only.

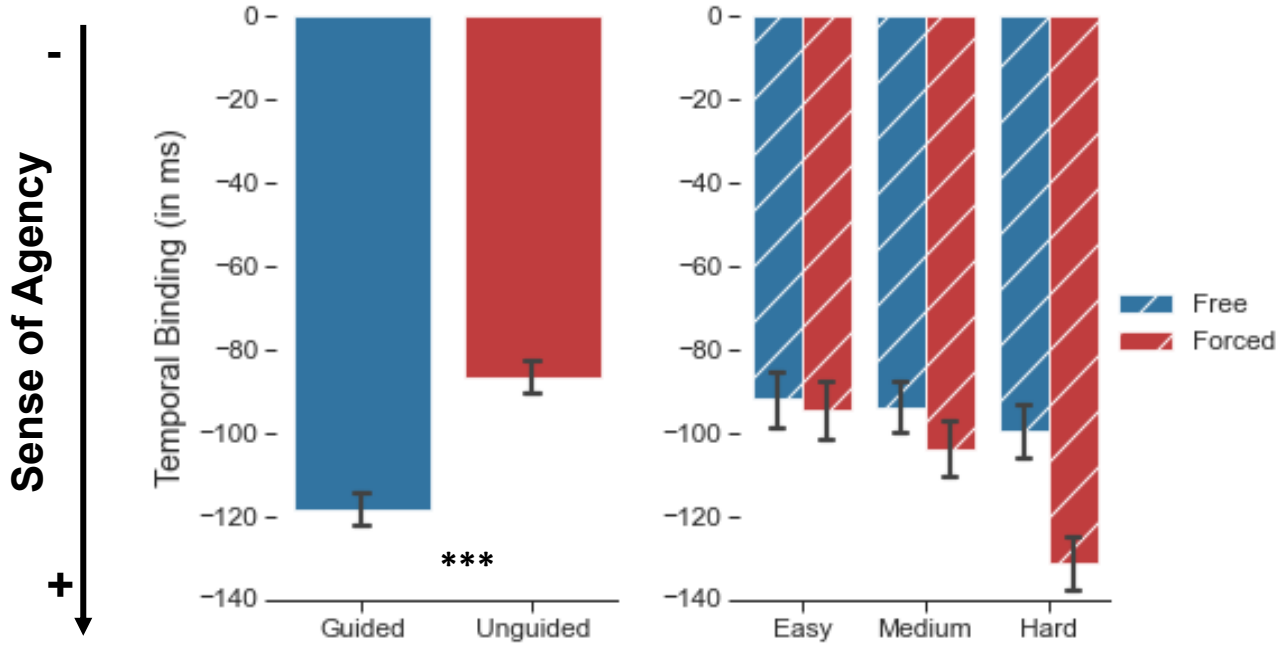

Figure 3. Mean TB associated with the two levels of the explicability factor (left plot), and with Levels of Automation (Free vs Forced) and Difficulty (Easy vs Medium vs Hard). Note that a high *negative* TB value is indicative of a high sense of agency. Error bars show +/−1 within-subject standard error of the mean (S.E.M). Three-stars: $p<0.001$.

3.3 Decision confidence

The repeated-measures ANOVA revealed significant main effects of explicability (guided: 6.42 ± 1.79; unguided: 6.30 ± 1.81; $F(1,36) = 5.87$; $p = 0.021$; $\eta_p^2 = 0.140$) and task difficulty (easy: 7.06 ± 1.28, medium: 6.44 ± 1.72, high: 5.57 ± 2.00; $F(2,72) = 113.86$; $p < 0.0001$; $\eta_p^2 = 0.760$), with participants showing greater confidence in their response in guided vs. unguided trials, and when the task was the easiest (**Figure 4**).

The explicability-by-automation-by-difficulty triple interaction was significant ($F(2,72) = 5.95$; $p = 0.005$; $\eta p^2 = 0.142$), as well as the explicability-by-automation ($F(1,36) = 31.58$; $p < 0.001$; $\eta p^2 = 0.467$) and the explicability-by-difficulty ($F(2,72) = 7.85$; $p = 0.003$; $\eta p^2 = 0.180$) interactions. The automation-by-difficulty interaction did not reach statistical significance ($F(2,72) = 2.51$; $p = 0.088$). Decomposition of the 3-way interaction using post hoc tests revealed that participants were least confident in trials with the least amount of supervision, i.e. when they had to make a free choice without receiving further information from the system (i.e. without the system indicating confidence in either direction; $p = 0.002$). Conversely, participants' confidence in their response was little affected by having to make a forced (as opposed to free) choice in trials where they received information from the system. Taken together, these results suggest that any additional information provided by the system to "guide" the participant's choice can support, if not restore, confidence in the response made, even in highly automated situations and when the task is most difficult (here, "hard" forced-choice trials). Confidence was significantly lower, except for the easiest trials ($p = 0.178$) in unguided free-choice trials than in unguided forced-choice trials (medium difficulty: $p = 0.014$; high difficulty: $p = 0.021$), whereas the difference between free- and forced-choice conditions was not significant in guided trials (all p's < 0.6).

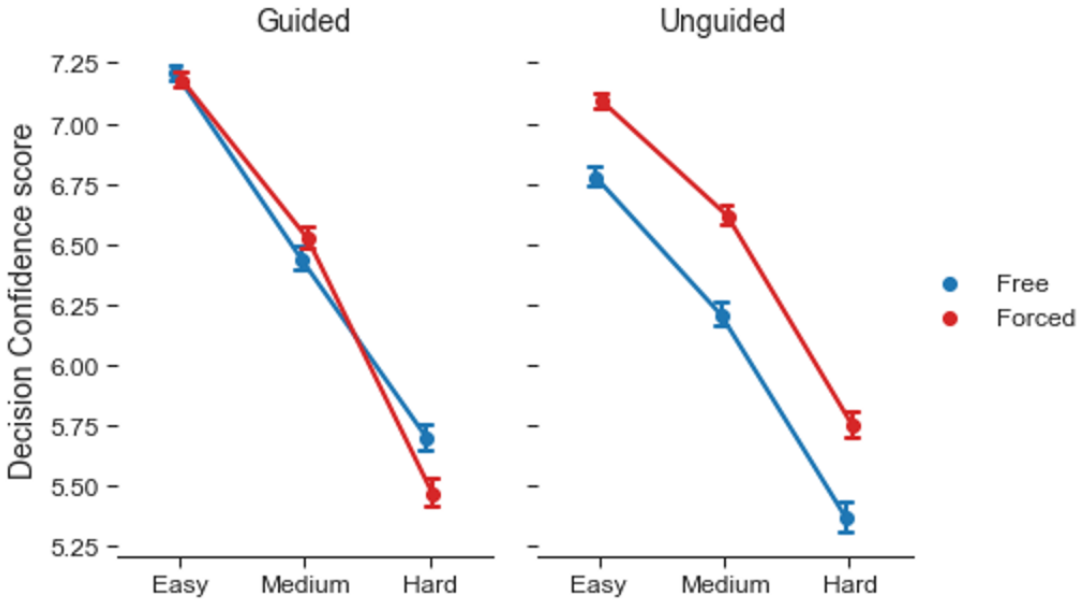

Figure 4. Average Decision Confidence scores associated with the Automation (Free vs Forced), Explicability (Guided vs Unguided) and Difficulty factors (Easy, Medium, High). Error bars show +/−1 within-subject standard error of the mean (S.E.M).

3.4 Acceptability Scales

As expected, system A was reported to be more usable than system B (65.12± 20.11 vs. 61.58 ± 22.23; paired t-tests, $p = 0.039$). In contrast, both systems had satisfaction scores that were not statistically different (64.60 ± 18. 78 vs. 64.55 ± 19.36; paired t-tests, $p = 0.969$) (**Figure 5**).

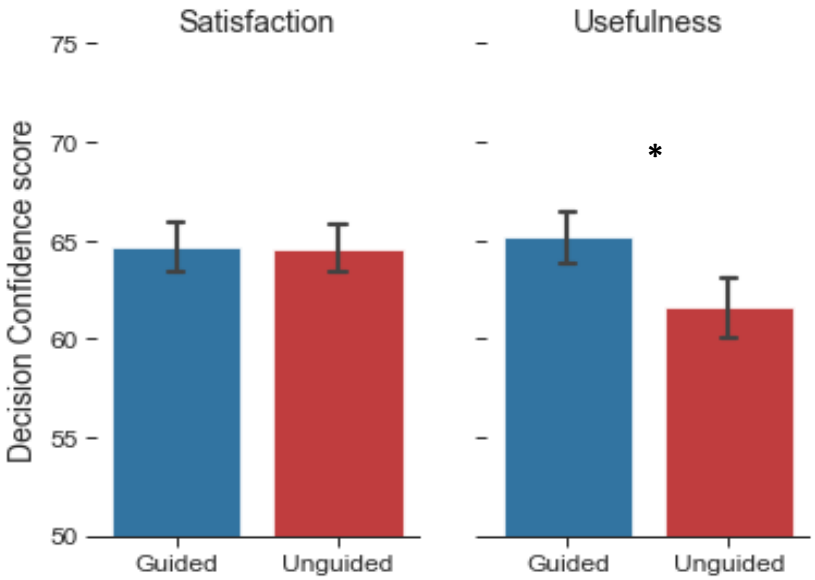

Figure 5. Average usefulness and satisfaction scores. Error bars show +/−1 within-subject standard error of the mean (S.E.M). One-star: $p < 0.05$.

3.5 Preliminary discussion – Experiment 1

Three main results were found. As predicted, increased explicability was associated with increased SoA, but also with better performance, greater decision confidence, and higher system acceptability. However, the relation between SoA and automation was not expected: the participants' SoA was greater in forced-choice than in free-choice trials, contrary to previous results (e.g. Barlas et al., 2019; Sidarus, Vuorre & Haggard, 2017a; Wenke, Fleming, & Haggard, 2010). Finally, task difficulty contributed differently to the measures of agency (TB) and decision confidence. Task difficulty was associated with a significant *decrease* in confidence, whereas the measure of agency *increased* with increasing task difficulty, but in forced-choice trials only. This last result was surprising because we would have expected that increasing the difficulty of the task would decrease the feeling of control experienced during the task.

One explanation for this result may lie in what is actually being assessed when using "temporal binding" as a measure of control, and more broadly in what is meant by the notion of "control" as operationalized in experimental tasks. “Control” can be framed in terms of *experience* (how much control is felt) or in terms of *resources* (how much control is used). Interestingly, understanding the notion of control in terms of experience or resources generates opposite patterns of relationship between task difficulty and SoA, just as we observe in our task. Indeed, the negative relationship between task difficulty and our implicit measure of agency is typical of the relationship previously found between task difficulty and a “resource-oriented” measure of control (Potts & Carlson, 2019).

To test this hypothesis more directly, we implemented the scales used by Potts and Carlson to disentangle reports of “control used” (CU) vs. reports of “control felt” (CF) (Potts et al., 2019). Whereas the CF scale measures the degree of control felt over the trial during the execution of the task, the CU scale quantifies the cognitive effort or resources invested in executing the task during each trial. We implemented these two types of explicit measures in the avoidance task used in Experiment 1, and predicted that each measure will exhibit different (positive vs. negative) relationships with task demands, in a way that is consistent with the patterns observed in the first experiment.

4 Experiment 2

4.1 Methods

4.1.1 Participant

Thirty-nine participants were recruited to participate in Experiment 2 (28 females, mean age: 34,23, SD: 9.24). Sample size calculation was based on the effects found in Experiment 1. A priori power calculation was performed using the G*Power software (Faul et al., 2009), with a power of 0.80 and two-sided alpha level set at 0.05. The number of participants required to detect a mean effect size of $d = 0.4$ in a paired comparison, with ~10% exclusion in the sample based on predefined exclusion criteria (see below), was thirty-nine. The inclusion/exclusion criteria were the same as in Experiment 1 (8 subjects were excluded on the basis of these criteria). The study was approved by the local ethics committee (reference: IRB n. 21-810) and all participants gave written informed consent before inclusion in the study, which was carried out in accordance with the declaration of Helsinki (1964, revised 2013). As for experiment 1, the study was conducted on the PRISME platform and was divided into two sessions, on two different days to reduce the effect of fatigue. Each session lasted 2 hours, and participants were paid 40 euros after completing both sessions.

4.2 General paradigm

The general procedure was similar to Experiment 1 (**Figure 1**), with the only exception of the explicit measure collected. Instead of measuring confidence, Experiment 2 measured the sense of control used/felt over the decision on each trial. Specifically, in alternating blocks and immediately after the temporal estimation, participants were asked to report either their sense of control used (presented as the amount of effort invested in performing the task) or control felt (presented as the degree of control felt while performing the task) using a Likert scale ranging from 1 (low control) to 8 (high control).

4.3 Data analysis

Percentage of correct responses and TB measures were analysed as in Experiment 1. The two measures of explicit control (control used and control felt) were analysed as a single dependent variable in a 2 × 2 × 2 × 3 repeated-measures ANOVA, with the level of automation (free vs. forced-choice), the level of explicability (guided vs. unguided), task difficulty (easy vs. medium vs. high) and the type of control measured (Used or Felt) as within-participants factors. Note that, for this ANOVA, we refrained from further interpreting the main effects because they conflate CF and CU

measures altogether. Our primary interest was to compare CF and CU measures with each other, and therefore the interpretation will focus on the interaction effects of the ANOVAs. The main effects are nevertheless reported in the text for the sake of completeness and transparency. The datasets used for the analyses are: "*All_data_Exp2.csv*", "*CorrectRates_Experiment2.csv*"and "Acceptability_Exp2.csv" (Vantrepotte et al., 2021)

5 Results

5.1 Avoidance task performance - % Correct Responses

The ANOVA revealed significant main effects of the explicability ($F(1,30) = 13,00$; $p = 0.001$; $\eta_p^2 = 0.302$) and the difficulty factors ($F(1,30) = 304,64$; $p < 0.001$; $\eta_p^2 = 0.910$) on mean correct response rates. Post hoc comparisons showed that the performance was higher in guided trials than in unguided trials ($p = 0.001$) and decreased with increasing difficulty (all p's < 0.0001). The explicability-by-difficulty interaction effect was not significant ($F(2,60) = 0.44$; $p = 0.645$).

5.2 Temporal Binding

We found significant main effects of the explicability ($F(1,30) = 11.06$; $p = 0.002$; $\eta_p^2 = 0.269$) and the automation factors ($F(1,30) = 8.21$; $p = 0.008$; $\eta_p^2 = 0.215$) with both the guided trials (**Figure 6, left panel**) and the forced trials (**Figure 6, right panel**) being associated with a stronger binding effect, i.e. shorter action-outcome intervals. The main effect of task difficulty was not significant ($F(2,60) = 0.23$; $p = 0.795$). No significant interaction effects were found (explicability-by-automation interaction: $F(1,30) = 0.50$; $p = 0.484$; explicability-by-difficulty interaction: $F(2,60) = 1.52$; $p = 0.484$; automation-by-difficulty: $F(2,60) = 0.56$; $p = 0.571$ – but see Suppl. fig. 2 for an analysis performed on aggregated data from both experiments). As for Experiment 1, the explicability-by-difficulty-by-automation triple interaction was not significant ($F(2,60) = 0.34$; $p = 0.7$).

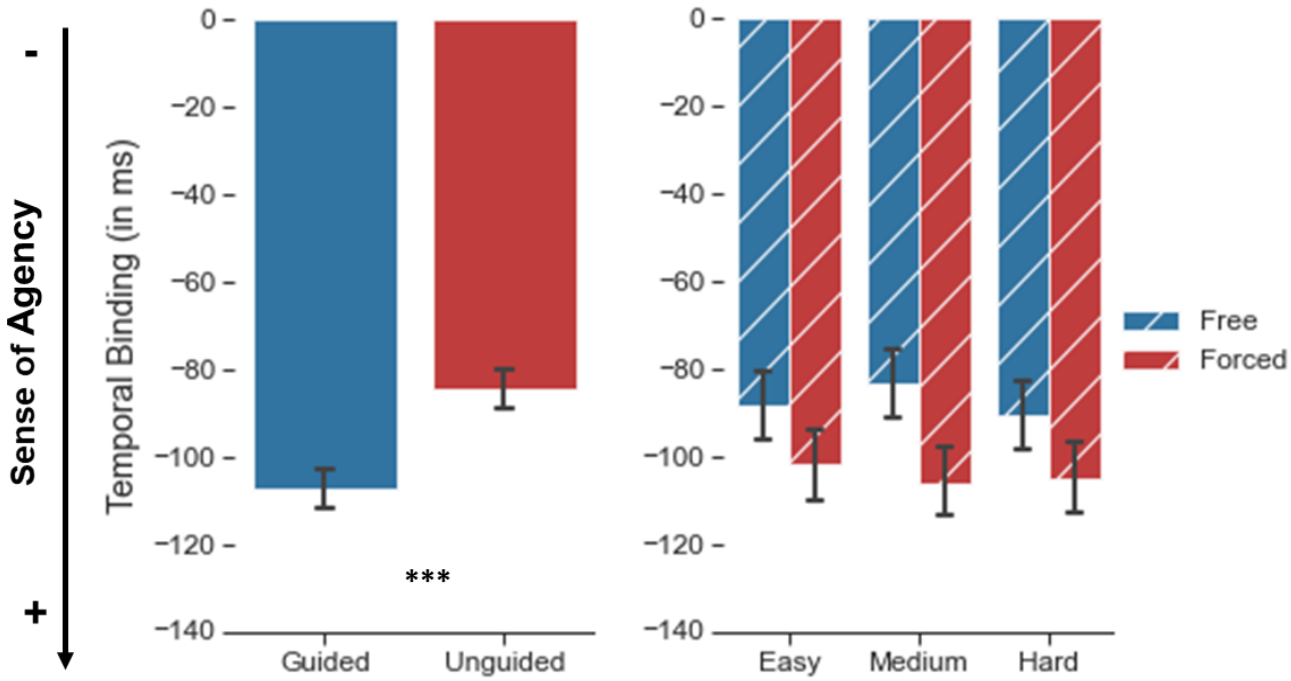

Figure 6. Mean TB associated with the two levels of the explicability factor (left plot), and with Levels of Automation (Free vs Forced) and Difficulty (Easy vs Medium vs Hard). Note that a high *negative* TB value is indicative of a high sense of agency. Error bars show +/−1 within-subject standard error of the mean (S.E.M). Three-stars: p<0.001.

5.3 Explicit measures of control

Significant main effects of automation (forced: 4.43 ± 2.13; free: 4.77 ± 2.02; F(1,30) = 9.88; p = 0.004 ; ηp^2 = 0.248), difficulty (easy: 4.46 ± 2.25, medium: 4.74 ± 1.99, high: 4.59 ± 1.99; F(2,72) = 6.45; p = 0.003 ; η_p^2 = 0.177) and type of control (CF: 5.07 ± 1.94; CU: 4.13 ± 2.12; F(1,30) = 18.52; p < 0.001 ; ηp^2 = 0.382) were found.

Participants tended to report experiencing higher control in guided, relative to unguided, trials, but this numerical difference did not reach significance (main effect of explicability: F(1,30) = 0.21; p = 0.653). Note that the automation-by-type of control interaction effect was not significant (F(2,60) = 1.57; p = 0.220), suggesting that the effect of automation (free > forced) was the same for both control measures (felt and used). The difficulty-by-automation interaction was significant (F(2,60) = 9.39; p < 0.001 ; ηp^2 = 0.238). The difficulty-by-type of control was significant (F(2,60) = 125.82; p < 0.001 ; ηp^2 = 0.807). As expected, the sense of control "used" (CU) increased, and the control "felt" (CF) decreased, with task difficulty (CU vs. CF on easy trials: p <0.001; medium trials: p <0.001); hard trials: p = 0.99). The explicability-by-type of control interaction was significant as well (F(2,60) = 7.11; p = 0.012 ; ηp^2 = 0.192, **Figure 7**). Although the explicability factor influenced the two control scores in opposite ways (CU: Guided < Unguided; CF: Guided > Unguided), post-

hoc analyses could not identify significant differences between guided and unguided trials when comparing CU and CF measures (guided, CU vs. CF: $p = 0.1$; unguided, CU vs. CF: $p = 0.3$).

Finally, we also found a significant automation-by-difficulty-by- type of control interaction effect ($F(2,60) = 9.39$; $p < 0.001$; $\eta p^2 = 0.238$). Post-hoc tests showed that participants reported more CU in free vs forced trials as difficulty increased (easy difficulty: $p = 0.99$; medium difficulty: $p = 0.025$; hard difficulty: $p < 0.001$, see **Figure 8, left panel**), while no difference was observed between free and forced trials, on any of the 3 levels of difficulty, for the CF measure (all p's > 0.9, see **Figure 8, right panel**). In sum, this triple interaction shows that the interplay between automation and difficulty was only discernible for the measure of control used, not for the measure of control felt. No other significant interaction effects were found.

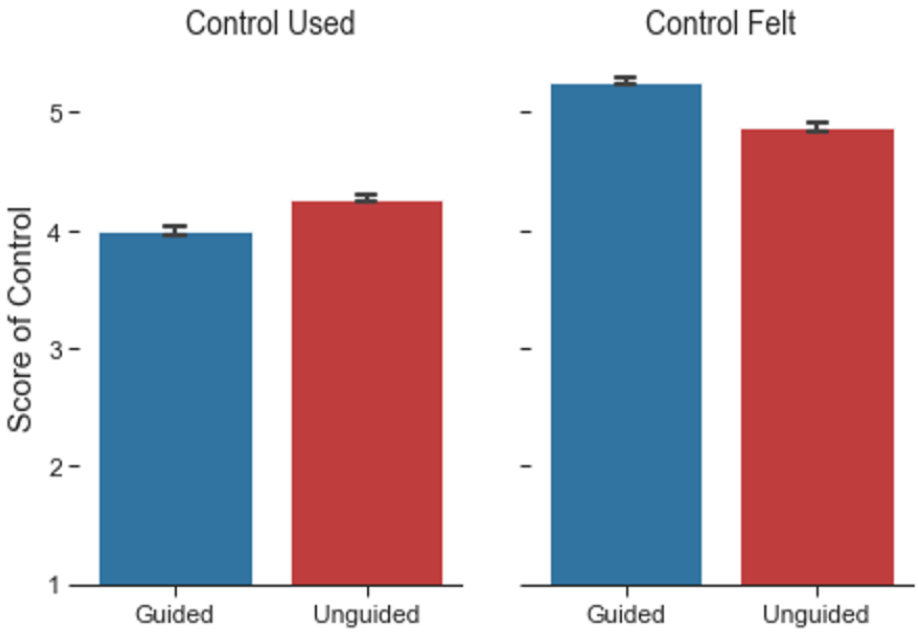

Figure 7. Reports of "Control Used" and "Control Felt" associated with the System Explicability (Guided vs Unguided) factor. Error bars show +/−1 within-subject standard error of the mean (S.E.M)

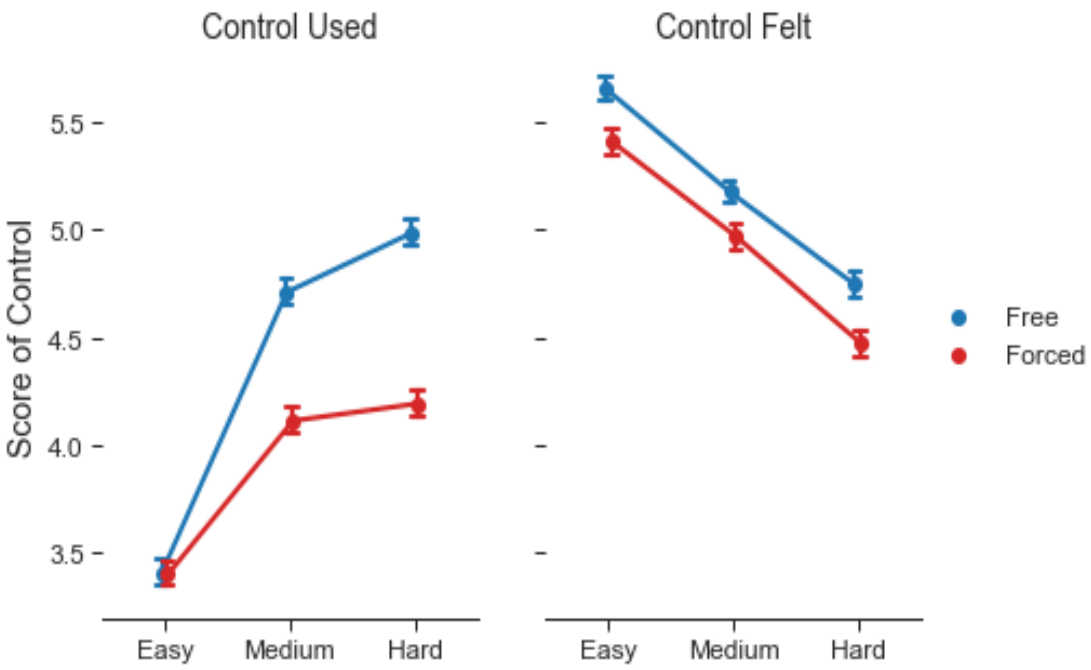

Figure 8. Reports of "Control Used" (left-hand chart) and "Control Felt" (right-hand chart) associated with the Automation (forced vs. free) and Difficulty (easy vs. medium vs. high) factors. Error bars show +/−1 within-subject standard error of the mean (S.E.M).

5.4 Acceptability Scales

No significant differences were found, whether in terms of system usefulness (system A vs. system B: 62.87± 22.70 vs. 61.20 ± 22.38; paired t-tests; p =0.43) or system satisfaction (63.51± 20.06 vs. 64.53± 19.58; paired t-tests; p = 0.53).

5.5 Preliminary discussion – Experiment 2

Three main results were found. Consistent with the results obtained in Experiment 1, increased explicability was associated with higher SoA for the implicit measure of control (temporal binding), but also with better performance. For the explicit measures of control, a significant interaction was found between type of control and explicability factors (**Figure 7**). Post-hoc tests directly comparing conditions for each control measure, however, were not significant (CU: Guided > Unguided, p = 0.3; CF: Guided < Unguided, p = 0.14). The unexpected relation between our implicit measure of agency (TB) and the level of automation was confirmed: the participants' SoA was greater in forced-choice than in free-choice trials. Finally, and as expected, different patterns of results were observed regarding the relation between difficulty and the subjective measures of control: control used

increased with difficulty, in a way that was consistent with the pattern observed for the TB measure, while control felt *decreased* with difficulty, in a way consistent with the pattern observed for confidence measures in Experiment 1. This result suggests, albeit indirectly, that temporal binding might be influenced by the cognitive resources engaged in the task (as measured by reports of "control used"). This relationship between cognitive resources and control experience is not new, although conflicting results have been reported (see Demanet et al., 2013; Damen et al., 2014; Howard, Edwards, & Bayliss, 2016; Sidarus et al., 2016). Further work is needed to determine whether the effect of task difficulty on TB reflects an influence of cognitive resources on the participant's experience, or whether it indicates that TB does not measure the experience of voluntary action per se (as suggested elsewhere, see, e.g., Hoerl et al., 2020) but rather is a marker of the amount of resources invested in a given task. In contrast, ease of action and performance would be more directly related to measures of “control felt” and confidence. We discuss these results in further detail in the next section.

6 General Discussion

Our two experiments aimed at characterizing the key factors responsible for the sense of agency (SoA) in an operator interacting with an automated system. As previously suggested, reducing the opacity of system decisions can contribute to improving human-machine interactions and, by extension, the operator's sense of control over the decisions made during the task (Norman, 1990; Berberian et al., 2012). To test this suggestion more directly, we designed a task in which levels of automation and explicability of systems were manipulated by implementing conditions in which the participant could or could not freely choose what to do (free vs. forced trials) and by providing or not providing the participant with the system's confidence in its decision (guided vs unguided trials). Finally, task difficulty was implemented as an additional factor that could modulate the hypothesized relationships between SoA, automation and explicability. Participants' choice and confidence in that choice, as well as participant's SoA and acceptability in the system decision, were collected throughout the task.

6.1 Increasing explicability to support human agency and confidence.

Our first objective was to determine whether making a system more "explainable," i.e., making it less opaque via communicating confidence in its decision, increased the participant's sense of agency in the task at hand. Analysis of our implicit measure of agency, i.e. temporal binding (TB), revealed an effect of explicability on SoA: guiding participants through the decision process led to a greater underestimation of the temporal interval between action and effect, a known marker of SoA in human participants (Haggard et al, 2002). This observation is consistent with previous experimental results, which show that providing subjects with information *before* they even take an action prospectively influences the feeling of control over the subsequent action (Wenke et al., 2010). This prospective effect of information on SoA would be due to its facilitating effect on the selection of the action – a facilitating effect that would be all the more important as action selection would be difficult to achieve (Chambon & Haggard, 2012; Voss et al., 2017; Jacquet et al., 2012; Jacquet et al., 2016). Interestingly, a similar prospective effect of prior information on participants' confidence in their decision has been reported (Rahnev et al., 2015; Fleming et al., 2016). In line with these results, we also found that participant's decision confidence (Experiment 1) and control felt (Experiment 2) both benefited from making the system more explainable (guided > unguided trials). In contrast, and as expected, our measure of "control used" showed a negative relationship with the explicability factor, with reported resources invested in the task being lowest in *guided* trials.

Taken together, these results demonstrate that the metacognitive information delivered by the system contributes to restoring the operator's SoA, and this increased SoA is also associated with greater acceptability of the system used. Thus, these results complement those obtained previously (le Goff et al, 2018) and demonstrate the role of system intelligibility on the SoA of operators interacting with automated systems.

6.2 Positive "Automation Surprise"

The automation factor was also found to alter the participants' sense of agency (SoA). Thus, in both Experiments 1 and 2, forced choice trials were associated with stronger temporal binding (TB), i.e. higher sense of agency, than free choice trials. These results positively relate SoA and *increased* automation, in stark contrast with previous studies also manipulating free and forced choice trials (Barlas et al., 2018, 2019; Caspar et al., 2016, 2018). These differences could be explained by (1) the nature of the action representations manipulated in our experiments, and by (2) the nature of the assistance provided by the system to the human operator.

First, even when "forced" to choose, participants retain proximal (i.e., motor) control over the decision insofar as they are the ones who physically press the button that implements this decision. As such, our "forced" trials differ from other experiments where proximal control is delegated to a robot (Engbert, Wohlschläger, and Haggard, 2008), to another agent (Sahaï, Pacherie, Grynszpan, & Berberian, 2019), or even to an external stimulation induced by transcranial magnetic stimulation (Haggard and Clark, 2003). The control experienced in forced-choice trials can also be explained by the fact that, very often, these forced trials are aligned with what the subject would have intentionally chosen. This alignment is illustrated by Suppl. Fig. 1 showing the proportion of responses as a function of confidence level. Compliance with system guidance is relatively high, suggesting that the system choice is (at least partially) aligned with the participant's choice.

Second, a key element to consider in explaining the positive impact of the automation variable on participants' SoA relates to the performance of the support system: in the present task, participants interacted with a system that was better than they were. This feature highlights two critical elements. First, following the system's suggestions leads to better performance than ignoring those suggestions and performing the task alone. It is now accepted that performance is a constituent of the feeling of control. (Metcalfe et Greene, 2007; van der Wel, Sebanz, & Knoblich, 2012; Wen et al., 2015; Ueda et al., 2021). As an illustration, Metcalfe and Greene (2007) found that people's self-agency judgment was correlated with their performance on the task, even when they were explicitly aware that their performance was largely the result of external factors. In a computer-assisted pointing task, Wen et al. (2015) also showed that performance plays a critical role in the

development of SoA, even when performance depends on decisions made by the system (assisted condition) rather than by the human operator himself (self-control condition). Similarly, our observation of increased SoA in the forced-choice condition could be due to better overall performance in this condition. In other words, during difficult trials, the system becomes a support for higher performance, which results in an increased sense of control by the participants.

While the increase in performance emphasizes the retrospective dimension of SoA, the assistance system could also have a *prospective* impact on SoA. As mentioned earlier, the system performs better overall than the participant. The system, due to its high level of performance, could be perceived by the participant as an effective way to reduce the intrinsic uncertainty of the task. Here, the increase in SoA could then be equated with an increased sense of "knowing" rather than an increased sense of "doing" (Koriat, 2000; Haggard & Chambon, 2012; Mylopoulos & Shepherd, 2020). Indeed, the increase in SoA on forced trials is primarily observed on difficult trials, where uncertainty is greatest for the operator in the absence of assistance. Another unexplored possibility is that participants may have gradually learned that complying with the highly confident system reduced uncertainty during decision- making and produced positive feedback on average. Belief in high performance can prospectively influence feelings of control over subsequent action outcome, which in our task would have resulted in stronger temporal binding on trials where the system was most confident. Finally, it should be noted that the proposed task is essentially a decision-making task. It is likely that here decision-related uncertainty is more important in the construction of participants' SoA than the sensorimotor cues classically emphasized (Synofzik et al., 2008). In the two tasks of the present study, SoA was indeed rather determined by post-hoc cues such as the valence of the outcome, and thus by the performance (of the system or the participant) that depends on the valence of this outcome. The results obtained on the confidence and acceptability measures confirm the influence of the system's assistance on the participants' subjective experience. In addition to experiencing greater control, participants also reported greater confidence in forced-choice trials compared to free-choice trials in both Experiments 1 and 2. Interestingly, this effect of automation on decision confidence (forced trials > free trials) disappeared when participants received prior information from the system (guided trials) (**Figure 4, left panel**), with participant expressing high confidence in *both* free- and forced-choice guided trials. Thus, as expected, the information provided by the system about the choice itself (forced choice) or confidence in the correct choice to make (guided trials) was beneficial to participants' confidence in their decision. Finally, acceptability scores showed that participants generally perceived the systems favorably, as cooperators rather than as interfering or malicious agents.

Our results highlight the link between automation, sense of agency and cognitive fatigue. In our task, the alignment between the intentions of the system and those of the operator should have the effect of decreasing the cognitive load and thus the fatigue of the participants, thus increasing their sense of agency. This positive effect of automation and explicability on SoA (see **Figure 8**, **left panel**) is consistent with other recent studies showing that sustained physical or cognitive effort alters agents' SoA. The reason for this alteration would be that the agency attribution mechanism is itself resource-consuming and that any effort invested in completing a task depletes the resources available for this attribution mechanism (Howard et al., 2016, but see also Minohara et al., 2016, showing that effort can *enhance* self-attribution under certain circumstances). This inference, however, remains highly hypothetical since no metrics of cognitive fatigue were collected in our protocol. Future work could focus on the specific challenges posed by cognitive fatigue in human-machine interactions (such as brain-machine interfaces, see e.g. Caspar et al., 2021). One promising line of research would be to test whether greater explicability and a greater sense of agency in the human partner can help mitigate the negative effects of cognitive fatigue on operator performance and system acceptability.

To summarize, we found that the participant's SoA increased as a function of the assistance provided by the automated system, whether this assistance was related to the decision itself (free vs. forced trials) or to the system's confidence in the best decision to make (guided vs. unguided trials). Note that these two types of assistance could interact during the task, with potentially different effects on participants' agentive experience. In forced-choice trials, complying with the system's choice can be seen as an effective way to reduce the intrinsic uncertainty of the task, especially when the task is difficult. The increase in participants' SoA in the forced-choice condition could therefore be driven by participants' (often verified) belief in high performance when following the system's advice. In contrast, in free-choice trials, where the subject was free to choose the action, communicating the system's confidence in the best action could ultimately have the effect of making action selection more fluid, and boosting subjects' sense of agency as a result. This effect of "selection fluency" on SoA is consistent with the previously observed effect on SoA of participants who received subliminal priming during action selection (Chambon & Haggard, 2012; Chambon & Haggard, 2013; Sidarus et al., 2013; Chambon et al. 2015). Another interesting interpretation of these results is that system guidance in free trials sometimes conflicts with participant choice. This conflict could lead to an experience of decisional “dysfluency" that would explain the reduction in SoA on free-choice trials compared to forced-choice trials. ~~Note, however, that participants perform very similarly in "guided" and "unguided" free trials, suggesting that guidance, while informing action~~

~~experience, has little influence on participants' choice.~~ The influence of task demands on temporal binding (TB) observed in Experiment 1 was not replicated in Experiment 2 (but see additional analysis on the aggregated data from both experiments, supp. Figure 3). One possible explanation is that Experiment 2, unlike Experiment 1, implemented an additional control measure explicitly probing the cognitive resources engaged in the task (the "control used" measure). This focus on the "resource" could have had the effect of helping the subject better regulate the resources invested in the task as a function of task difficulty, with the consequence of leveling TB across difficulty levels.

7 Conclusions and perspectives

In two experiments, we showed that explicability could be used as a lever to improve the agency of operators interacting with automated systems. Improving the explicability of the decisions of the system itself increases – in free-choice trials – or restores – in forced-choice trials – operators' sense of agency (SoA). Importantly, we found that the difficulty of the task at hand modulated the relationship between explicability and automation. When the subject acts alone and receives no assistance, the uncertainty associated with the difficulty of the task impairs the formation of a reliable SoA. When the subject is assisted and the assistance system is well calibrated, both in terms of objective performance and metacognitive evaluation, the subject's SoA naturally benefits from any cues provided by the system. It remains to be determined whether, and to what extent, operators' agency is affected when the assistance is less reliable, in terms of performance and/or metacognitive calibration. Answering this question would require selectively manipulating either the performance level of the assistance system or the calibration of the confidence returned by the system, and observing the impact of this manipulation on SoA and acceptability.

We did not find a direct link between explicit (e.g. "control felt") or implicit (Temporal Binding) measures of SoA (see Supplementary Material, Correlation analyses). One reason for the lack of direct linkage could be due to the nature of the measures themselves, and the fact that implicit and explicit measures diverge when task demands vary (Dewey & Knoblich, 2014; Obhi & Hall; 2011). With respect to the implicit measure, the action-effect interval estimation method used in this experiment cannot determine whether the observed interval reduction ("binding") is due to a shift in the perception of action onset or outcome onset. It is not therefore possible to decide whether increased explicability resulted in greater anticipation of the outcome (the outcome is temporally shifted toward the action) or delayed awareness of the action (the action is temporally shifted toward the outcome). Implementing a version of the measure that collects both the perceived onset of the action and the perceived onset of the subsequent outcome (with a Libet clock-type measure; Libet,

2002) could allow for a finer-grained assessment of the impact of the different variables tested (explicability, automation, difficulty) on the perception of action-outcome relationships, which form the basis of the sense of agency.

In conclusion, our study shows that providing some level of assistance to users can improve their performance without negatively affecting their sense of control (Coyle et al, 2012; Berberian, 2019; Ueda et al, 2021). A positive effect of automation was indeed found in this experiment, and this effect was reinforced by various factors such as the sharing of a common intention, but also the increased reliability and explicability of the system. Making the system less opaque can also reduce physical and cognitive distance between the system and the operator, a distance that can be detrimental to both performance and user experience in human-machine interactions (Sheehan and Sosna, 1991). Our results further demonstrate the potential value of studying the subjective experience of control in human-machine interaction as a guideline for making the systems we create more agentic and acceptable (Le Goff et al, 2018).

Author contributions

Q.V., B.B., M.P. and V.C. developed the study concept. Testing, data collection and data analysis were performed by Q.V. Q.V. drafted the manuscript. B.B. and V.C. provided critical revisions. All authors approved the final version of the manuscript for submission.

Declaration of Competing Interest

None.

Supplementary material

Datasets related to this article can be found at https://osf.io/cme7d/, an open-source online data repository hosted at Open Science Framework (Vantrepotte et al., 2021).

Acknowledgements

This work was supported by the Agence Nationale de la Recherche (ANR) grants ANR-17-EURE-0017 (Frontiers in Cognition), ANR-10-IDEX-0001-02 PSL (program 'Investissements d'Avenir') and ANR-16-CE37-0012-01 (ANR JCJ), ANR-19-CE37-0014-01 and ANR-21-CE37-0020-02 (ANR PRC). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

Balsdon, T., Wyart, V., Mamassian, P., 2020. Confidence controls perceptual evidence accumulation. *Nature Communications*, 1–11.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085.

Barlas, Z. (2019). When robots tell you what to do: Sense of agency in human- and robot-guided actions. *Consciousness and Cognition*, *75*, 102819.

Barlas, Z., Hockley, W. E., & Obhi, S. S. (2018). Effects of free choice and outcome valence on the sense of agency: Evidence from measures of intentional binding and feelings of control. *Experimental Brain Research*, *236*(1), 129–139.

Berberian, B. (2019). Man-Machine teaming: A problem of Agency. *IFAC-PapersOnLine*, *51*(34), 118–123.

Berberian, Bruno, 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.

Bigenwald A. & Chambon V. (2019). Criminal responsibility and Neuroscience: No Revolution Yet. Frontiers in Psychology: Theoretical and Philosophical Psychology, 10, 1406.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Caspar, E. A. (2017). Coercition et perte d'agentivité. *médecine/sciences*, *33*(5), 543–547.

Caspar, E. A., Cleeremans, A., & Haggard, P. (2018). Only giving orders? An experimental study of the sense of agency when giving or receiving commands. *PLOS ONE*, *13*(9), e0204027.

Caspar, E. A., De Beir, A., Lauwers, G., Cleeremans, A., & Vanderborght, B. (2021). How using brain-machine interfaces influences the human sense of agency. *Plos one*, 16(1), e0245191.

Chambon, V., Filevich, E., & Haggard, P. (2014). What is the human sense of agency, and is it Metacognitive?. In *The cognitive neuroscience of metacognition* (pp. 321-342). Springer, Berlin, Heidelberg

Chambon, V., & Haggard, P. (2012). Sense of control depends on fluency of action selection, not motor performance. *Cognition*, *125*(3), 441-451.

Chambon, V., & Haggard, P. (2013). 14 Premotor or Ideomotor: How Does the Experience of Action Come About? *Action science: Foundations of an emerging discipline*, 359

Chambon, Valérian, Moore, J. W., & Haggard, P. (2015). TMS stimulation over the inferior parietal cortex disrupts prospective sense of agency. *Brain Structure and Function*, *220*(6), 3627–3639.

Chambon, Valérian, Sidarus, N., & Haggard, P. (2014). From action intentions to action effects: How does the sense of agency come about? *Frontiers in Human Neuroscience*, *8*, 320.

Chambon, Valérian, Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067–1079.

Chambon, Valerian, Wenke, D., Fleming, S. M., Prinz, W., & Haggard, P. (2013). An Online Neural Substrate for Sense of Agency. *Cerebral Cortex*, *23*(5), 1031–1037.

Christoffersen, K., & Woods, D. D. (2002). 1. How to make automated systems team players. In *Advances in human performance and cognitive engineering research*. Emerald Group Publishing Limited.

Coyle, D., Moore, J., Kristensson, P. O., Fletcher, P., & Blackwell, A. (2012). *I did that! Measuring users' experience of agency in their own actions*. 2025–2034.

[dataset] Vantrepotte, Q., Berberian, B., & Chambon, V., *Leveraging Human agency data*, 2017, Open Science Framework, Version 1 , https://osf.io/cme7d/

Dewey, J. A., & Knoblich, G. (2014). Do implicit and explicit measures of the sense of agency measure the same thing? *PloS One*, *9*(10), e110118.

Di Costa, S., Théro, H., Chambon, V., & Haggard, P. (2018). Try and try again: Post-error boost of an implicit measure of agency. *Quarterly Journal of Experimental Psychology*, *71*(7), 1584–1595.

Ebert, J. P., & Wegner, D. M. (2010). Time warp: Authorship shapes the perceived timing of actions and events. *Consciousness and cognition*, 19(1), 481-489.

Engbert, K., Wohlschläger, A., & Haggard, P. (2008). Who is causing what? The sense of agency is relational and efferent-triggered. *Cognition*, 107(2), 693-704.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*.

Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2016). Metacognition about the past and future: Quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, *2016*(1).

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, *23*(8), 931–939.

Haggard, P., & Chambon, V. (2012). Sense of agency. *Current Biology*, *22*(10), R390–R392.

Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and cognition*, 12(4), 695-707.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382–385.

Haggard, P., & Eitam, B. (2015). *The Sense of Agency*. Oxford University Press.

Haggard, P., & Tsakiris, M. (2009). The Experience of Agency: Feelings, Judgments, and Responsibility. *Current Directions in Psychological Science*, *18*(4), 242–246.

Hoerl, C., Lorimer, S., McCormack, T., Lagnado, D. A., Blakey, E., Tecwyn, E. C., & Buehner, M. J. (2020). Temporal binding, causation, and agency: Developing a new theoretical framework. *Cognitive Science*, 44(5), e12843.

Howard, E. E., Edwards, S. G., & Bayliss, A. P. (2016). Physical and mental effort disrupts the implicit sense of agency. *Cognition*, 157, 114-125.

Jacquet, P. O., Chambon, V., Borghi, A. M., & Tessari, A. (2012). Object affordances tune observers' prior expectations about tool-use behaviors. *PloS one*, *7*(6), e39629.

Jacquet, P. O., Roy, A. C., Chambon, V., Borghi, A. M., Salemme, R., Farnè, A., & Reilly, K. T. (2016). Changing ideas about others' intentions: updating prior expectations tunes activity in the human motor system. *Scientific reports*, *6*(1), 1-15.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

Klien, G., Woods, D., Bradshaw, J., Hoffman, R., & Feltovich, P. J. (2004). Ten Challenges for Making Automation a 'Team Player' in Joint Human-Agent Activity. *Intelligent Systems, IEEE*, *19*, 91–95.

Koriat, A. (2000). The Feeling of Knowing: Some Metatheoretical Implications for Consciousness and Control. *Consciousness and Cognition*, *9*(2), 149–171.

Lausic, D. (2009). *Communicating Effectively: Exploring Verbal and Nonverbal Behaviors and How They Affect Team Coordination*.

Lawrence, M. A., & Lawrence, M. M. A. (2016). Package 'ez'. *R Package Version*, *4*(0).

Le Bars, S., Devaux, A., Nevidal, T., Chambon, V., & Pacherie, E. (2020). Agents' pivotality and reward fairness modulate sense of agency in cooperative joint action. *Cognition*, *195*, 104117

Le Bars, S., Bourgeois-Gironde, S., Wyart, V., Sari, I., Pacherie, E., & Chambon, V. (2020, December 8). Motor coordination and strategic cooperation in joint action. https://doi.org/10.31234/osf.io/xbm34

Le Goff, K., Rey, A., Haggard, P., Oullier, O., & Berberian, B. (2018). Agency modulates interactions with automation technologies. *Ergonomics*, *61*(9), 1282–1297.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.

Libet, B. (2002). *The timing of mental events: Libet's experimental findings and their implications*.

Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology: General*, *136*(2), 184–199.

Minohara, R., Wen, W., Hamasaki, S., Maeda, T., Kato, M., Yamakawa, H., ... & Asama, H. (2016). Strength of intentional effort enhances the sense of agency. *Frontiers in psychology*, 7, 1165.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: A review. *Consciousness and Cognition*, *21*(1), 546–561.

Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Philosophical Transactions of the Royal Society B*, *376*(1828), 20200052.

Mylopoulos, M., & Shepherd, J. (2020). The experience of agency. In *The Oxford Handbook of the Philosophy of Consciousness*.

Norman, D. A. (1990). THE PROBLEM OF AUTOMATION: INAPPROPRIATE FEEDBACK AND INTERACTION, NOT OVER-AUTOMATION. *Philosophical Transactions of the Royal Society of London*, 18.

Obhi, S. S., & Hall, P. (2011). Sense of agency and intentional binding in joint action. *Experimental Brain Research*, *211*(3–4), 655.

Pacherie, E. (2013). Intentional joint agency: Shared intention lite. *Synthese*, *190*(10), 1817–1839.

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Poizat, G., Bourbousson, J., Saury, J., & Sève, C. (2009). Analysis of contextual information sharing during table tennis matches: An empirical study of coordination in sports. *International Journal of Sport and Exercise Psychology*, *7*(4), 465–487.

Potts, C. A., & Carlson, R. A. (2019). Control used and control felt: Two sides of the agency coin. *Attention, Perception, & Psychophysics*, *81*(7), 2304–2319.

PP, R. (1964). Human experimentation. Code of ethics of the world medical association. Declaration of Helsinki. *British Medical Journal*, *2*(5402), 177–177.

R Core Team. (2017). *R: The R Project for Statistical Computing*. https://www.r-project.org/

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science*, *26*(11), 1664–1680.

Sahaï, A., Desantis, A., Grynszpan, O., Pacherie, E., & Berberian, B. (2019). Action co-representation and the sense of agency during a joint Simon task: Comparing human and machine co-agents. Consciousness and Cognition, 67, 44-55.

Sheehan, J. J., & Sosna, M. (1991). *The boundaries of humanity: Humans, animals, machines*. University of California Press.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYSTEMS LAB.

Sidarus, N., Chambon, V., & Haggard, P. (2013). Priming of actions increases sense of control over unexpected outcomes. *Consciousness and cognition*, *22*(4), 1403-1411

Sidarus, N., Palminteri, S., & Chambon, V. (2019). Cost-benefit trade-offs in decision-making and learning. *PLoS computational biology*, *15*(9), e1007326

Sidarus, N., Vuorre, M., & Haggard, P. (2017). How action selection influences the sense of agency: An ERP study. *NeuroImage*, *150*, 1–13.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239.

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 353-382). Springer, Boston, MA.

Ueda, S., Nakashima, R., & Kumada, T. (2021). Influence of levels of automation on the sense of agency during continuous action. *Scientific Reports*, *11*(1), 1–13.

Van Der Laan, J. D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies*, *5*(1), 1–10.

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2012). The sense of agency during skill learning in individuals and dyads. *Consciousness and cognition*, 21(3), 1267-1279.

Vogel, D. H., Jording, M., Esser, C., Weiss, P. H., & Vogeley, K. (2021). Temporal binding is enhanced in social contexts. *Psychonomic Bulletin & Review*, 1-11.

Voss, M., Chambon, V., Wenke, D., Kühn, S., & Haggard, P. (2017). In and out of control: brain mechanisms linking fluency of action selection to self-agency in patients with schizophrenia. *Brain*, *140*(8), 2226-2239

Wen, W., Yamashita, A., & Asama, H. (2015). The influence of action-outcome delay and arousal on sense of agency and the intentional binding effect. *Consciousness and Cognition*, *36*, 87–95.

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition*, *115*(1), 26–38.

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA*, *310*(20), 2191–2194.