

Consequential sound induces illusory distortions in the perception and prediction of robot motion

Joel Currie^{1*}, Maria Elena Giannaccini² and Patric Bach¹

¹School of Psychology, University of Aberdeen, St Machar Drive, Aberdeen, AB24 3FX, UK.

²School of Engineering, University of Aberdeen, Elphinstone Rd, Aberdeen, AB24 3UE, UK.

*Corresponding author(s). E-mail(s): j.currie.22@abdn.ac.uk;

Contributing authors: elena.giannaccini@abdn.ac.uk; patric.bach@abdn.ac.uk;

Abstract

For efficient human-robot interaction, human operators need to be able to efficiently represent the robot’s movements in space and predict its next steps. However, according to frameworks of Bayesian multisensory integration, features outside the motion itself – like the consequential sounds a robot makes while it moves – should affect how otherwise identical motions are perceived. Here, we translate an established psychophysical task from experimental psychology to a human-robot interaction context, and can measure these distortions to motion perception. In two series of preregistered studies, participants watched a humanoid robot make (forward and backward) reaching movements. When the robot hand suddenly disappeared, they reported its last seen location, either with the mouse cursor (Experiment 1a and 1b) or by matching it to probe stimuli in different locations (Experiment 2a and 2b). The results revealed that even small changes to the robot’s consequential sound robustly affect participants’ visuospatial representation of its motions, so that the motion appeared to extend further in space when accompanied by slightly (100 ms) longer sounds compared to slightly shorter sounds (100 ms shorter). Moreover, these sound changes do not only affect where people currently locate the robot’s motion, but where they anticipate its future steps. These findings show that sound design is an effective medium for manipulating how people represent otherwise identical robot actions and coordinate its interactions with it. The study acts as proof of concept that psychophysical tasks provide a promising tool to measure how design parameters influence the perception and prediction of robot motion.

Keywords: human-robot interaction, social robotics, representational momentum, movement sonification, cue-integration, motion perception.

1 Introduction

Robots are increasingly used in close interaction with humans. In manufacturing, robots work together with humans on flexible production lines for tasks that cannot be fully automated

[1, 35, 64]. Close physical human-robot interaction is important in rehabilitation [37] to support patients in everyday life and to improve their agency [48, 52], as well as in therapy and socially demanding service work [28, 78]. Robots are also used for teleoperation in environments that

are not safe for human operators [61], or where domain experts are not available locally [10].

What limits productive human-robot interactions in these and other applications is that robots often behave in a way that people find unnatural [66]. Human-to-human cooperation is built on the ability to effortlessly “read” and predict the behaviour of one’s interaction partners [5, 63]. From simple tasks like handing over tools or moving a table to more complex ones such as assembling an object together, efficient cooperation relies on the ability to represent our partner’s actions [68]. These interactions are supported by sophisticated internal models that capture how intentional (human) interaction partners behave within a given environment [4, 15, 27, 43]. These models do not only supply observers with higher-level semantic information about which action another agent is currently executing [63], but also with more fundamental visuospatial information about these actions’ lower-level kinematics, such as changes in spatial location, direction, and speed, which are crucial for guiding moment-to-moment interactions in a shared task space [54]. Importantly, observers represent these kinematic parameters in a predictive manner, capturing not only the movement’s current state but also how it will likely develop in the immediate future, so that one’s own actions can be directed not towards where the motion is now but where it will be in the future [4, 27, 43]. Non-biological agents like robots do not fit these human-centric internal models [75], however, making it difficult for people to understand and predict their behaviour [13, 42].

Different approaches have investigated how human understanding of robot behaviour can be improved, from making robots communicate via gestures [36], endowing them with human-like social cues such as eye-gaze [60], to giving them the ability for speech [40], with limited success [6]. Here, we investigate an important but typically neglected factor that may affect the observer’s perception of robot kinematics: the consequential sound the robot produces as it moves – the noise of its motors, the friction of its moving parts and the whirring of fans [70]. Human motion perception is multisensory. The visual perception of a moving stimulus is automatically integrated

with information received from other senses, and prior information about how the motion is likely to continue, to arrive at a statistically optimal estimate of what is observed [18, 74, 76]. The consequential sounds that accompany robot motions provide a central component to these estimations. Sounds should therefore be involuntarily integrated into the perception of its motion, and – especially when sound and motion are not well aligned – should induce illusory distortions in how the motion is represented. Identical visual motions should therefore be perceived differently depending on the features of the sounds that accompany them.

So far, only a handful of studies have investigated how consequential sound is integrated into the perception of robot motion, and these studies have measured exclusively its impact on the evaluation of the robot’s higher-level socio-emotive features, such as its attractiveness, quality of movement and perceived safety [62]. Nothing is known about how sound changes the perceptual representation of robot motion itself. Yet, as noted above, investigating such visuospatial changes is crucial because these features are what is ultimately used by human interaction partners to plan their own responses within the shared task space. Distortions to a movement’s visuospatial representation will therefore have a direct effect on the actions that human operators direct towards the robot. Moreover, even higher-level socio-emotive responses to robot movements likely stem in part from lower-level representations of its kinematic features (e.g., its jerkiness, smoothness, etc.), or from mismatches of motion signals in different channels (“uncanniness” [38, 65]). So far, however, no tasks are available that would enable researchers to probe how human observers visuospatially represent a robot’s motion.

The goal of the current study is to (1) develop an experimental task that can sensitively measure the predictive kinematic representations that humans derive of motions and (2) make it usable for HRI research. We will use (3) the robot’s consequential sound as a testing bed to investigate whether these representations can be effectively measured and manipulated. We draw upon the well-established family of representational

momentum-like tasks [14, 22], which ask participants to localise the last seen position of briefly seen motions (e.g., with a mouse pointer, or by comparing it to “probe” stimuli presented in different locations on the screen). Findings from this task have shown that, to localize moving objects, humans do not just rely on the parameters of the motion itself but improve their estimates by drawing on all available evidence across the multisensory perceptual sphere, as well as their prior expectations about how the motion will continue, and integrate it following the principles of Bayesian computation [11, 31]. The resulting representations of observed motions therefore do not veridically represent the visual motion input, but are distorted towards the motion’s expected next steps, which allows agents to compensate for inherent delays in motor control, making real-time interactions with a dynamic environment possible [19, 20, 54]. Research has shown, for example, that location judgments of moving objects are (erroneously) displaced away from the objects true location towards the expected next steps of the motion sequence, and that these misperceptions are driven by one’s prior knowledge of the forces acting on the object [21], intuitive knowledge of physics [16], context, and in social interactions, and the goals attributed to an agent [4, 24–27, 43, 44]. Importantly, there is also prior evidence that accompanying sounds are integrated with these predictive overestimations, making the motion appear more or less pronounced depending on the duration of the sound that accompanies it [71].

Here, we use this task to probe people’s kinematic representation of robot movements, and measure how it is affected by consequential sound. In four studies, participants viewed brief sequences of a robot hand in side view reaching or withdrawing (see [24] for similar design in human action perception). The robot’s hand disappeared at an unpredictable point during the action, and participants reported its last seen location, either with their mouse cursor (Experiments 1a and 1b) or by comparing it to static probe stimuli showing the robot hand displaced forwards or backwards in time (Experiments 2a and 2b). To test whether the robot’s consequential sound affects motion representation, we manipulated the duration of the sound that accompanied the

movements, as an important ecologically valid sound component. As a complex system made up of various components, a robot’s consequential sounds will not always begin and end simultaneously, due to inherent delays and their position in the operational hierarchy. For example, a motor may be most active during the early stage of a motion, while a fan that cools it may continue to operate even after the motion has ceased. If sound duration is integrated into the perception of visual motion, then changes to sound duration should affect the extent of the perceived motions. Reaches and withdrawals should therefore be reported to have travelled further along their trajectory when accompanied by a longer sound than a shorter sound [71].

Our research strategy is as follows. In Experiment 1a and the preregistered Experiment 1b, we ask participants to localise the motion disappearance points using a computer mouse. These studies show indeed that the consequential sound is integrated with human perception of robot action and affects how otherwise identical motions are visuospatially perceived. Importantly, mouse responses rely on the same visuospatial motor maps that people use to coordinate the movements of their own limbs within a dynamic environment [30, 50, 51] and can therefore serve as a proxy for the visuospatial processes guiding dynamic human-robot-interaction. However, while they robustly capture the effect of sound on spatial localisation, they are subject to various biases that render them unable to establish how perceptual representations more generally relate to the objective motion that was perceived, specifically whether the representation of robot movement is predictive or lags behind the perceived motion. To provide this crucial test, in Experiments 2a and 2b we ask participants to make their motion disappearance judgments not with a mouse cursor, but by comparing them to static “probe” stimuli presented after. The non-spatial nature of these responses eliminates the confounding biases and shows for the first time that human representation of robot motions is inherently predictive, capturing not only its last seen location but being enriched by expectations of how it will develop in the future.

2 Experiments 1a and 1b

2.1 Methodology

2.1.1 Participants

Participants (Experiment 1a: 51 participants, 35 female, mean age 21.2, SD = 3.32, 41 right-handed; Experiment 1b: 51 participants, 19 female, mean age 32.6 years, SD = 10.13, 48 right-handed), were recruited using the University of Aberdeen’s research participation scheme (Experiment 1a) and Testable Minds (<https://minds.testable.org/>) (Experiment 1b). They gave electronic informed consent as part of the experiment briefing and were reimbursed with course credits or £4.70, each experiment took approximately half an hour.

In Experiment 1a, the final sample of 44 participants provides .80 power to detect effect sizes of $d = .43$ ($n_p^2 = 0.159$). For Experiment 1b, the final sample of 42 gives us .80 power to detect effects of at least $d = .44$ ($n_p^2 = 0.165$). Prior studies investigating multisensory integration in biological motion outside HRI [24–26] revealed that effect sizes in similar paradigms are consistently of this size or larger ($d = .52$ to $d = 1.23$).

2.1.2 Apparatus

Visual stimuli were filmed on a Canon m50 mirrorless camera and edited using Lightworks (<https://lwks.com/>) and VLMC (<https://www.videolan.org/vlmc/>). Audio stimuli were created using samples from <https://freesound.org/> which were edited using Audacity [3] and MATLAB [41]. The experiment was programmed using the Inquisit platform [46] in an online format. Participants used their own personal devices to complete the experiment. Included devices were personal computers with a mouse and keyboard. Participants with mobile devices were asked not to participate and excluded if they did. Participants were asked to wear headphones during the experiment, and their ability to hear the sounds played was self-verified on the basis of test sounds (simple beeps) at the start of the experiment.

2.1.3 Stimuli

Visual Stimuli:

Visual stimuli showed the side profile of the

humanoid robot Pepper (Aldebaran) [55], completing a reaching (leftward), or a withdrawing (rightward) action. The grasping trajectory was designed to emulate a human reaching towards, and then grasping, an object. To reduce unnecessary uncertainty for participants unfamiliar with robot kinematics the designed grasp/withdraw trajectory was biologically inspired. As humans sample human motion daily, they have an inherent robust model of human motion [2, 57]. Due to this it was important for the robot’s morphology to match the human-like trajectory profile. Of the robots available to our lab, the Pepper robot best suited this task. Background details in the video were replaced with a uniform black background using the video editing software Lightworks.

Experiment 1a: All stimuli were derived from an initial set of three reach actions, starting on the right side of the screen, reaching towards the left, consisting of 32 frames in total. Six sequences of motion were generated from this original set, each 5 frames long (6 including the initial still frame), three of them showing reaches and three showing withdrawals. The action sequences started with a neutral frame, which was randomly chosen between three alternatives (frame 13, 14, 16). Reaches and withdrawals were generated by either stepping forwards (e.g., frame 15-17-19-21-23) for reaches, and backwards (e.g., frame 11-9-7-5-3) for withdrawals through the original stimulus sequences. Due to the three different starting positions, each action sequence could terminate in different locations in space, either closer to the centre of the screen, in a medium distance from the centre, or further out in the periphery. Each frame of the sequence was displayed for 82 ms, and immediately replaced with the next frame without gap.

Experiment 1b: As in Experiment 1a, three reach and three withdrawal sequences were generated from the original reach sequences. In contrast to Experiment 1a, the starting position of the action sequence remained consistent across sequences (frame 14), while the extent of the motion could vary between 3, 4, or 5 frames (not including the initially presented still starting frame). As before, reaches were created by stepping forward through the initial sequence and withdrawals by stepping backwards through

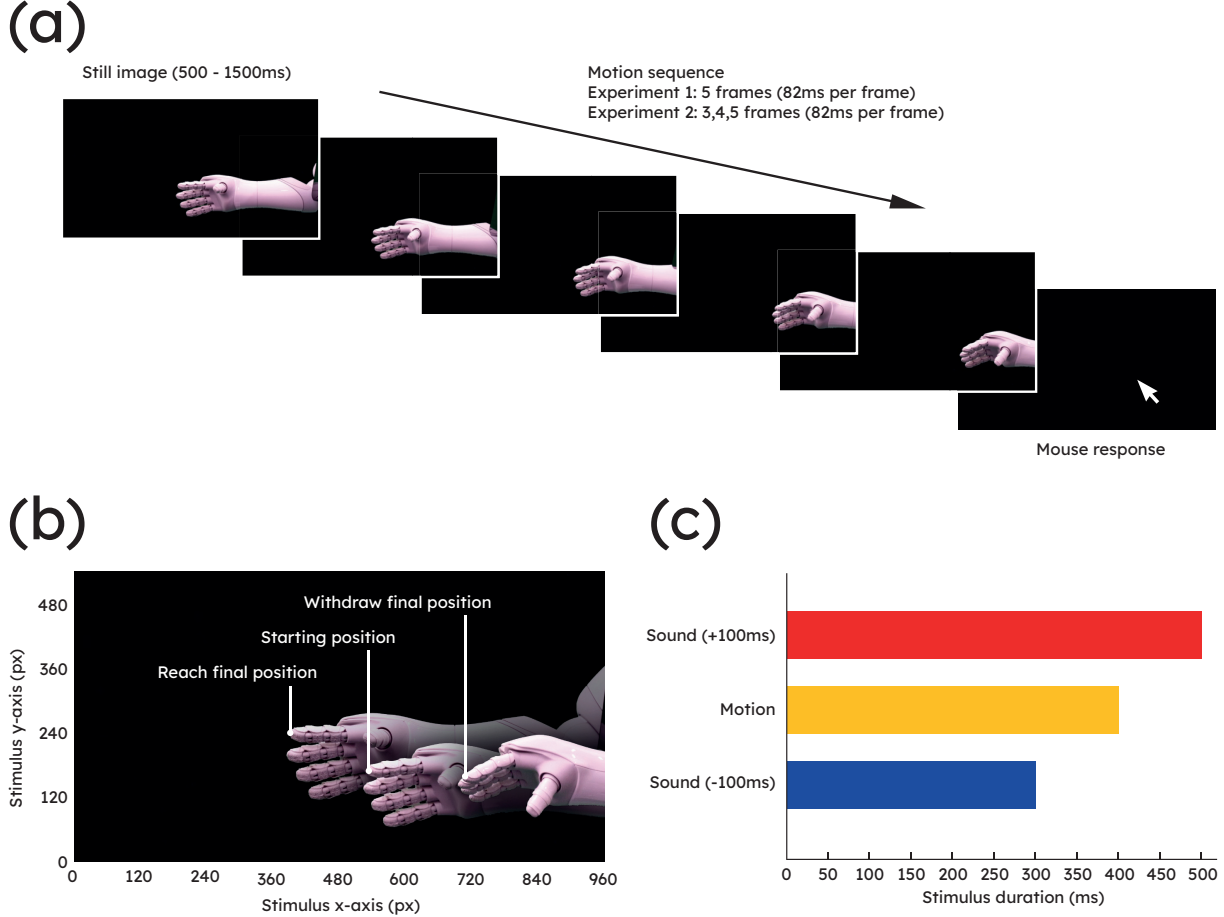


Fig. 1: A: Trial sequence example and relative lengths of presented visual and auditory stimulus. This trial sequence shows a withdraw. During the response stimulus the participant uses their mouse to indicate the last perceived location of the robot's index finger. B: Starting position and final Reach and Withdraw positions from a 5-frame motion. C: Relative motion and auditory offsets

them. Due to the different sequence lengths, each action sequence could again terminate in different locations in space, either closer to the centre of the screen, in a medium distance from the centre, or further in the periphery. Each frame within the sequence was displayed for 82 ms and immediately replaced with the next frame without gap.

Auditory Stimuli

Audio stimuli were designed to emulate the consequential noise a robot produces while moving. They were generated using two recorded samples of consumer motors from <https://freesound.org/>. The frequency of the samples was lowered to give the impression of larger motors (as seen in a robot such as Pepper). Dead zones in the audio

were removed. Experiment 1a: Two versions of the audio stimuli were generated, with durations of 310 ms and 510 ms, representing an offset of (-100 ms and + 100 ms) relative to the visual motion sequences. These offset values were chosen from literature [71], to induce the most detectable shifts in perception.

Experiment 1b: Six versions of the designed audio were generated. Three represented a 100 ms offset after the termination of the motion sequences, and three a 100 ms offset before their offset. This resulted in durations of 146 ms and 346 ms for the 3-frame sequence, 228 ms and 428 ms for the 4-frame sequence, and 310 ms and 510 ms for the 5-frame sequence.

2.1.4 Procedure

The participants first gave informed consent and received experimental instruction. They then proceeded to complete eight training trials, identical to the experimental trials. Training trials could be repeated if the participants felt they did not fully understand the task. Both Experiment 1a and 1b consisted of 120 trials, presented in 2 blocks of 60 trials each, with conditions following in pseudo-randomized order, counterbalanced so that each combination of the three sequence lengths/s-starting positions, two action directions, and two sound durations (100 ms longer or shorter than the movement durations) were repeated an equal number of times. In Experiment 1b, after the 120 experimental trials were completed, participants were asked to make a (free text) guess of the experimental hypothesis, to test whether the effects obtained can be explained by participants guessing the experimental hypothesis and adapting their behaviour accordingly (i.e., demand effects). Additionally, participants were asked questions probing their awareness of a change in the two sound conditions during the experiment (for details see in the supplementary information, testing for demand effects). The experiment lasted about 30 minutes in total.

At the beginning of each trial, participants were presented with a static image of the first frame of the movement. Following a randomly generated delay between 1000 ms and 2500 ms, the action sequence (reach or withdrawal) and auditory stimuli were presented. Action sequence and auditory stimuli always began synchronously after the random interval so that participants could form a causal relationship between the sound they were hearing and the robot hand moving [71]. The sequences stepped forward (for reaches) or backward (for withdrawals) through the stimulus sequence in five frames (Experiment 1a) or three, four or five frames in Experiment 1b. The last frame was immediately replaced by a black screen. The auditory stimuli stopped playing either 100 ms before the action sequence offset, or 100 ms after. The participants were asked to use their mouse and click to accurately indicate the last seen location of the tip of the robot’s index finger. They had five seconds maximum to respond. The next trial started after they clicked a green

marker in the centre of the screen, so that their mouse was centred for the start of the next trial.

2.1.5 Analysis

To measure the localisation error in participant responses, the displacement between the real coordinates of the robot’s index finger (x_t, y_t), and the mouse response coordinates of participants (x_r, y_r) were calculated. Since participants completed the experiment on their own personal devices, device resolutions were captured (x_{dr} and y_{dr}) and used to scale participants’ individual responses to a universal pixel size (1920×1080). Localisation error on the x-axis and y-axis are calculated using Equations 1 and 2, respectively.

$$e_x = \left(\frac{1920}{x_{dr}} x_r \right) - x_t \quad (1)$$

$$e_y = \left(\frac{1080}{y_{dr}} y_r \right) - y_t \quad (2)$$

Zero values on both axes indicate a perfect match of presented and reported disappearance points. Positive values of e_x denote a rightward shift in responses relative to the actual final position of the robot’s hand, and negative values represent leftward responses. Positive and negative values of e_y represent an upwards and downwards shift in responses, respectively.

The dataset was pre-processed and analysed using the statistical computing language R, Version 4.2.2 [58]. This included the exclusion of participants and trials (Section 2.1.6), and calculating the standardized localisation error for individual participants for each experimental condition.

Statistical analysis was completed using the function ‘ezANOVA’ from the package ‘ez’ [34] for the standardized localisation error, indicating the difference between the real location of the robot’s index finger’s disappearance point and the location the participants identified with their mouse click. Localisation error on the x-axis (e_x) was measured as the displacement between participants’ response coordinate (x_r), and the actual termination x-axis coordinate of the robot’s hand (x_t). Likewise on the y-axis, localisation error (e_y) is represented as the difference between participants’ response coordinate (y_r) and the real

termination y-axis coordinate of the robot’s hand (y_t).

2.1.6 Exclusion criteria

Participants were excluded if they used a mobile device without mouse. This was the case for four participants in Experiment 1a, and five in Experiment 1b. Additionally, participants with an aggregate mean localisation error (displacement between target stimulus and participants response across all conditions) greater than 10% of displayed stimulus size were excluded. From Experiment 1a three participants were excluded based on this criterion, and one from Experiment 1b. In Experiment 1b, three participants were excluded because they consistently responded after the allocated response time interval of 5 seconds.

In Experiment 1a, trials with localisation error greater than 3 SD from the median were excluded. As preregistered (https://aspredicted.org/blind.php?x=G6Z_XD7) in Experiment 1b, trials with localisation error greater than 3 SD from the median were excluded, as well as trials with response times shorter than 200 ms, or longer than 3 SD from the sample median ($\approx 2.2\%$).

2.2 Results - Experiments 1a and 1b

Each participant’s mean e_x localisation error were entered into $2 \times 2 \times 3$ repeated measures analyses of variance (ANOVA), with factors Action Direction (Reach vs. Withdrawal), Sound (-100 ms vs. +100 ms) and End Position of motion termination relative to the centre of the participant’s screen on the x-axis (Centre, Middle, Outer). Data was analysed using the same ANOVA model for both Experiment 1a and 1b.

Our predictions primarily concern the localisation errors on the x-axis, as the majority of the robot’s grasp motion is a translation on the x-axis, with limited variance on the y-axis (see Figure 1). Our main hypothesis is tested by the interaction of Sound and Action Direction, indicating that longer sounds bias responses more strongly leftwards for (leftwards moving) reaches and more rightwards for (rightwards moving) withdrawals than shorter sounds. While

not directly relevant to our hypotheses, and statistically independent from it, pilot work with this paradigm has also revealed an interaction between Action Direction and End Position. This reflects an increasing “pull” of participants’ location judgments towards the centre of the screen the more peripherally the motions terminate, in line with the well-known tendency for visual judgments to stabilize towards average disappearance locations across trials [39, 56].

As we have no further predictions, all other main effects and interactions should be treated as incidental unless meeting a (Bonferroni-adjusted) alpha threshold of .006 to account for hidden multiplicity in an ANOVA. [9]. Analysis of the y-axis data is presented in the Supplementary Information (see Experiment 1, y-axis results), and generally replicates all findings on the x-axis.

2.2.1 Experiment 1a

Analysis of participants’ localisation errors revealed the predicted two-way interaction of Action Direction and Sound ($F(1, 44) = 12.1, p = .002, n_p^2 = 0.216$). As can be seen in Figure 2, people reported the disappearance point of reaches (leftward motions) more leftward when the robot’s consequential sound extended beyond the motion than when it terminated before it, and more rightwards for withdrawals (rightward motions). As hypothesised, longer sounds therefore bias participants’ responses in the direction of the robot’s motion more strongly than shorter sounds, even though exactly the same visual motions were shown.

The ANOVA model also indicated a significant intercept ($F(1, 44) = 136, p < .001, n_p^2 = 0.75$). As apparent from Figure 2, participants’ localisation errors were generally positive, indicating that they reported the disappearance point of the robot’s index finger systematically more rightwards than it really was. This rightwards bias is statistically independent of our hypothesis tests and common for stimuli that extend leftwards such as ours, as localisation responses are typically attracted by the stimulus’ centre of mass (e.g., the middle of the hand, [8, 24, 27, 43, 44]).

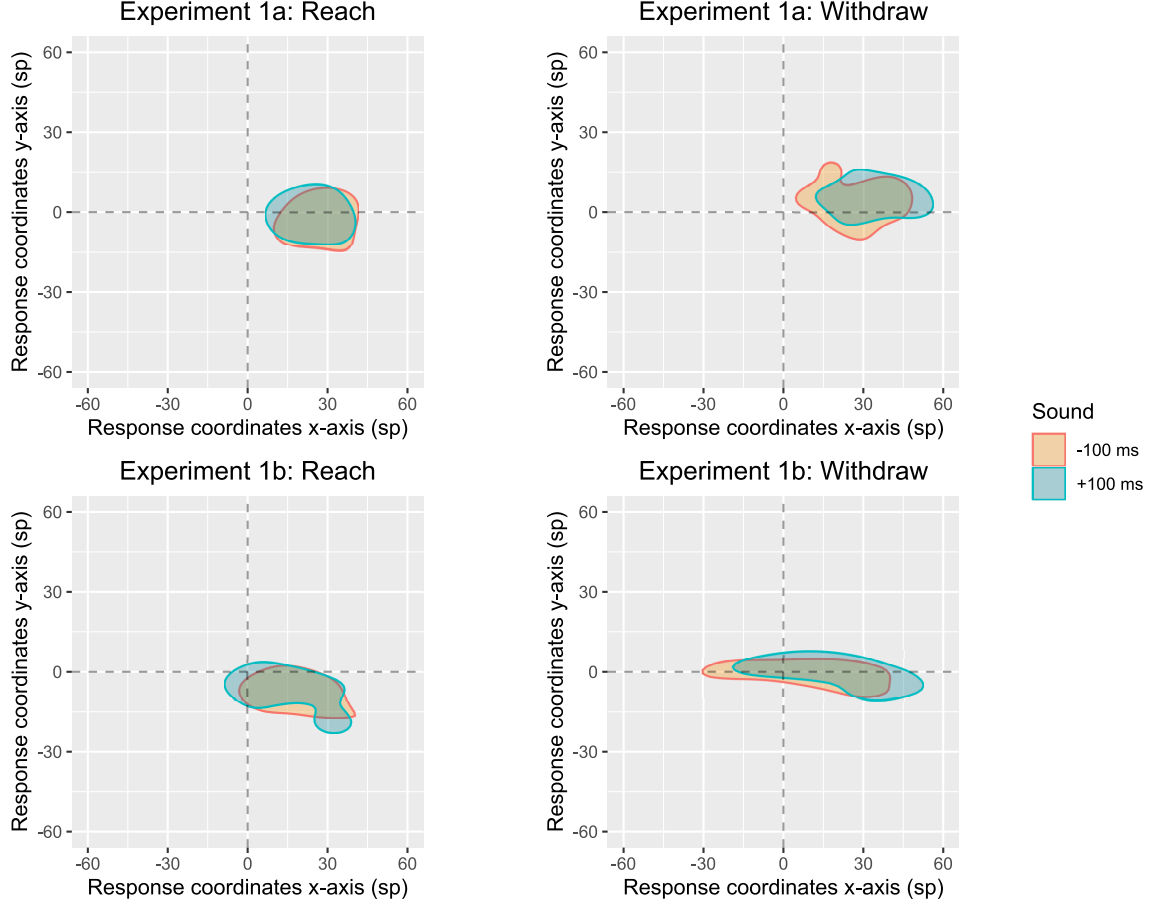


Fig. 2: 2D kernel density estimation for spatial distribution of response coordinates expressed as the difference between the real final coordinate of the robot’s index finger and participants’ response coordinate on the x-axis and y-axis. Response coordinates are shown in universal Scaled Pixels (SP). The (0,0) coordinate represents real final position on any given trial (0 SP difference on each axis). The (0,0) point for reaches has been placed above that for withdrawals as the real final position for reaches was above that of withdrawals

Finally, the expected interaction between Action Direction and End Position was found ($F(2, 88) = 50.72, p < .001, n_p^2 = .61$), as well as a main effect of End Position ($F(2, 88) = 9.64, p < .001, n_p^2 = 0.179$). These findings reflect the predicted general bias of mouse localisations towards the perceptual history, which average towards the centre of the screen [39, 56]. Our ANOVA models indicated no further main effects or interactions that met the adjusted threshold (all $F < 1.21, p > .328$).

2.2.2 Experiment 1b

Experiment 1b fully replicated the results of Experiment 1a. The analysis revealed, first, the predicted (and preregistered) two-way interaction of Action Direction and Sound, ($F(1, 41) = 71.9, p < .001, n_p^2 = .637$). As in Experiment 1a, people reported the disappearance of leftward going reaches more leftwards when the sound extended beyond the motion than when it terminated before it, and more rightwards for withdrawals (Figure 2, lower panel).

Like in Experiment 1a, our model also replicated the general biases affecting mouse localisation responses, which are independent of our hypotheses. It revealed a significant intercept, ($F(1, 41) = 17.9, p < .001, n_p^2 = 0.304$), replicating the general rightwards bias of localisation responses towards the robot hand’s centre of mass. Moreover, it replicated the general bias of mouse responses towards the centre of the screen, as indicated by an interaction between Action Direction and End Position, ($F(1, 82) = 5.79, p = .004, n_p^2 = 0.124$) as well as a main effect of Action Direction ($F(1, 41) = 17.9, p < .001, n_p^2 = 0.304$). Thus, as in Experiment 1a, participants’ localisations did not cover the whole of the motion but were biased centrally, and this inward bias increased the further outwards the motion terminated.

Our ANOVA models indicated no further main effects or interactions that met the adjusted threshold (all $F_s < 4.83, p > 0.011$), with the exception of a main effect of Sound ($F(1, 41) = 12.2, p = .002, n_p^2 = 0.231$), representing that participants responded further rightwards when presented with longer sounds, as opposed to shorter sounds, which is independent of our hypotheses. It is most likely due to the fact that the effect of sound offset on motion over-estimation was more pronounced for rightwards moving reaches (See Figure 2, lower panel), therefore creating an overall bias towards the right.

2.2.3 Between-experiment comparison

In an exploratory post-hoc analysis, we verified whether the localisation biases in the motion direction that were induced by longer sounds were larger in Experiment 1b than in Experiment 1a, due to the increased uncertainty about the end-positions and timing of the motion. We therefore ran the same ANOVA model on the pooled data of Experiment 1a and 1b, with Experiment (1a vs. 1b) added as an additional between-participants factor. This indeed revealed a three-way interaction between Action Direction, Sound and Experiment, ($F(1, 85) = 10.64, p = .002, n_p^2 = .111$), revealing that the sound had a larger effect on the perceived motion extent in Experiment 1b than Experiment 1a.

3 Experiments 2a and 2b

Experiments 1a and 1b provided the first evidence that the consequential sound of a robot influences how people represent the low-level visuospatial characteristics of its motion. When participants localised the disappearance point of briefly presented robot movements, their judgments were systematically influenced by the sounds it produced, so that the motions appeared to extend further into space when accompanied by a longer sound and not as far when accompanied by a shorter sound, particularly if the motion endpoints could be less reliably predicted in Experiment 1b than in Experiment 1a.

The mouse movements used for spatial localization are assumed to draw upon similar visuospatial motor maps [30, 50, 51] that are used to coordinate the movements of their own limbs within a dynamic environment and therefore capture processes during robot-human interaction within a shared task space. However, localisation responses are subject to well-known biases that we also observe here, such as a rightwards bias towards the hand’s centre of mass [8, 24, 27, 43, 44], and bias towards the middle of the screen, the starting point of the mouse pointer [39, 56] and the average location of judgments made before [12]. While these biases are independent of our hypotheses, they make it impossible to assess how participants’ representation of the observed motions more generally relates to their actual kinematics. A classic finding is that the human representation of observed motion is predictive, so that people misperceive motions displaced into the future, extrapolated towards the motion’s expected next steps [14, 18, 20]. These predictive displacements are assumed to be a major component of people’s ability to interact with a dynamic (social) environment [4, 67], compensating for delays in motor control [50, 51, 53, 54] and allowing our own responses to be planned towards where actions will terminate in the future rather than where they are right now.

To probe this predictive representation of observed robot motion, Experiment 2a and 2b changed how the influence of sounds on motion perception is measured. Participants saw and

heard the same stimuli as in Experiment 1b. However, now, they judged the motion disappearance points not by moving the mouse, but by comparing the hand’s last seen location to a static “probe” comparison image, which was presented shortly after the motion disappeared and showed the robot’s hand either further along the motion sequence than it really was (either +1 or +3 frames forwards) or not as far (either -1 or -3 frames backwards). Participants simply indicated, through a press of a button, whether the presented probe image was identical or different than the last seen image in the motion.

Probe tasks are well validated in motion perception research. They have been used to measure the motion of both naturalistic [24, 25, 27] (Hudson 2016 papers; Hudson 2017) and abstract stimuli [23], and how they are affected by sound [71]. In contrast to mouse responses, they more directly probe the perceptual representation of moving stimuli, without drawing upon visuospatial “motor” maps that are used to spatially guide one’s limbs to targets in the environment [30, 50, 51]. Importantly, the non-spatial nature of the required button presses renders them unaffected by the biases acting on mouse judgments and allows us to directly measure the predictive component of robot motion perception in Experiments 1a and 1b. We should therefore find that participants are more likely to identify probe stimuli further along the trajectory with the hand’s last seen location, compared to probe stimuli in a previous part of the motion. Moreover, if sounds are integrated with the perception of the motion, these mis-localisations should increase when the motions are accompanied by longer sounds than when accompanied by shorter sounds.

3.1 Methodology

3.1.1 Participants

Participants in Experiment 2a were recruited through the University of Aberdeen’s research participation scheme (24 participants, 19 female, mean age 24.1, SD = 6.91, 21 right-handed). In Experiment 2b, recruitment occurred through the participant recruitment platform Testable minds (32 participants, 22 male, mean age 31.9 years, SD = 8.09, 29 right-handed.). Participants

provided electronic informed consent as part of the experimental briefing and were reimbursed with course credits or £7.20 respectively. Each experiment took approximately 45 minutes.

The final sample of Experiment 2a, 19, provides .9 power to detect effect sizes of $d = .787$ ($n_p^2 = 0.357$). For Experiment 2b, the final sample of 26 provides .95 power to detect effect sizes of $d = .736$ ($n_p^2 = 0.360$).

3.1.2 Procedure

No adaptations were made to the apparatus or stimuli (sections 2.1.3) in the transition from Experiments 1b to Experiments 2a and 2b.

Participants first gave informed consent and received experimental instruction. Afterwards, eight training trials (identical to the experimental trials) were completed. These training trials, and the instructions preceding then, could be repeated if the participant wished. The experiment consisted of 288 trials, presented in six blocks of 48 trials each. Within each block, conditions were presented randomly, and counterbalanced such that each combination of the two action directions, two sound offsets and three action lengths were presented equal number of times. Between each experimental block there were two-minute breaks, with a countdown indicating the rest time remaining, and re-displaying the experimental instructions. At the end of the experiment, as in Experiment 1b, participants were asked to make a free-text guess of the experimental hypothesis, and were given a funnel debrief, investigating their ability to detect the sound manipulation (see in the supplementary information, testing for demand effects).

As in Experiment 1a and 1b, at the start of each trial participants were presented with a static image of a robot’s hand. Following a randomly generated delay between 1000 ms and 2500 ms the robot would either reach to the left, or withdraw to the right. Synchronously, the robot’s consequential sound would start to play as it moved across the participants’ screen. As in Experiment 1b, the robot stepped through the action sequence by three, four or five frames. The last frame was immediately replaced by a black screen. The

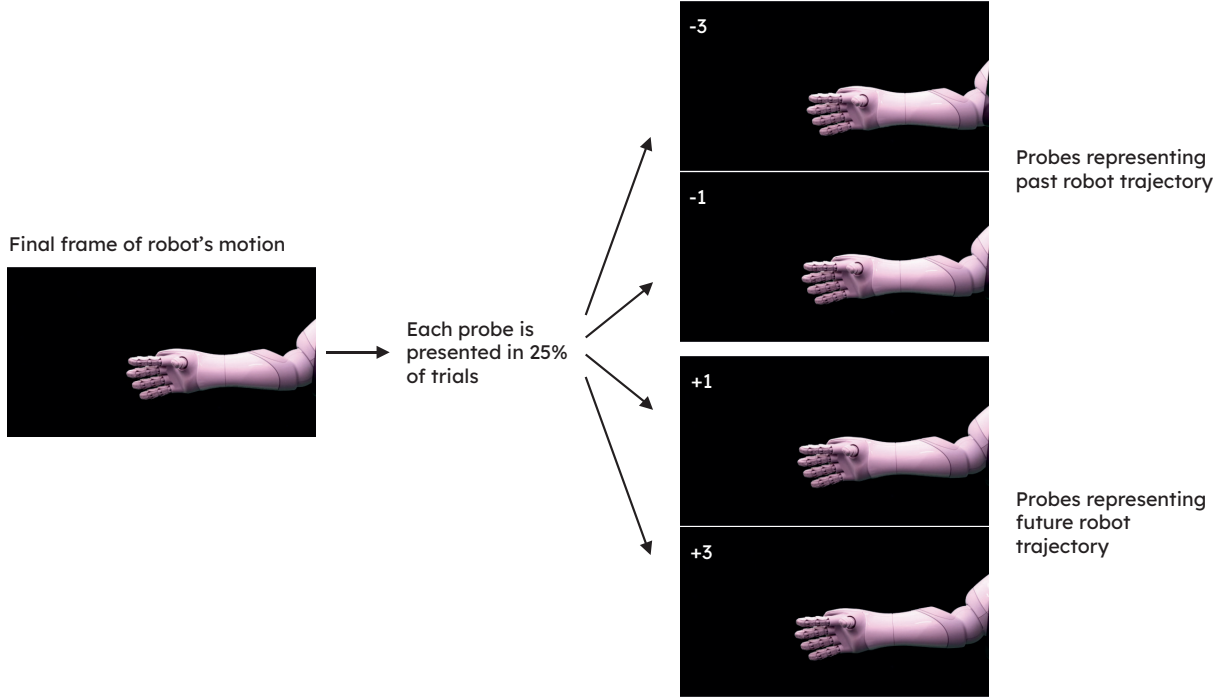


Fig. 3: Possible static probes that participants selected to be the 'same' or 'different' to the final frame of the action sequence

robot's consequential sound either stopped 100 ms before the robot finished moving, or 100 ms after its motion terminated. After a blank screen of 430 ms, participants were then presented with a static probe image of the robot's motion and asked to judge if the displayed probe was 'different' to the last position of the robot's hand. The presented static probe could show the hand either three frames behind its final position, one frame behind, one frame ahead, or three frames ahead, relative to the robot's direction of motion (see Figure 3).

'Different' responses were recorded by the participant pressing the 'spacebar' of their personal device. Participants were asked to signal "same" responses – that they perceived the presented probe hand to be in the same position as the hand's last position before the black screen – by not pressing a key. In Experiment 2b, auditory feedback (a simple chime) was given when the participants pressed the spacebar, to let them know that their response had been recorded. Participants had a maximum of 5 s to make their judgement. The next trial began 5 s after the probe response stimulus was displayed, independent of participant response.

3.1.3 Analysis

The dataset was pre-processed using R [58]. Data from participants who did not finish the experiment were excluded as well as the data from participants who met one of the exclusion criteria (see 3.1.4). Motion Overestimation, expressed as the weighted means of the proportion of participants 'same' response, was calculated for each participant for each condition. Statistical analysis was completed using the 'ezANOVA' function from the 'ez' [34] library.

While shifts in perception using similar 'probe tasks' has been quantified in various ways (see [22], for overview), we used the weighted means approach [17], as this technique provides a measure that weights forward and backwards probes (future/past) with equal importance, while weighting responses that are further in the past, or future of the action sequence more strongly. The weighted means of the proportion of 'same' responses per condition was calculated as is represented in Equation 3.

$$p = \begin{pmatrix} -3 \\ -1 \\ 1 \\ 3 \end{pmatrix}, MO_w = \frac{\sum_{p=-3}^{p=3} \left(\frac{\bar{x}_{s,p}}{\bar{x}_{T,p}} \right) \times p}{\sum \frac{\bar{x}_s}{\bar{x}_T}} \quad (3)$$

In which MO_w is the weighted shift in motion overestimation, and positive MO_w indicates forward displacement from the robot’s hand in the direction of motion. Negative MO_w represents displacement against the direction of motion. p denotes the probe image presented to participants, with options $(-3, -1, 1, 3)$, corresponding to the probe image being three frames into the past of the robot’s action sequence, one frame in the past, one frame in the future, or three frames into the future. $\bar{x}_{s,p}$ represents the sample mean of ‘same’ responses at a single probe, while \bar{x}_s , the total mean of same responses over all probes. $\bar{x}_{T,p}$ denotes the mean of the total responses (‘same’ plus ‘different’) at a single probe, and \bar{x}_T the total responses across all probes.

3.1.4 Exclusion Criteria

Participants were excluded if their mean proportion of ‘same’ responses was less than 10% or greater than 90%. This was the case for zero participants in Experiment 2a and Experiment 2b. Additionally, participants were excluded if they didn’t make at least 10% more ‘different’ responses for outer probes $(-3, +3)$ than inner probes $(-1, +1)$, suggesting an insensitivity to the probe differences, or general lack of attention. This was the case for 5 participants in Experiment 2a, and 6 participants in Experiment 2b. The final participant pool had the following demographics: Experiment 2a: 19 participants, 16 female, mean age 23.8, SD = 6.43, 16 right-handed; Experiment 2b: 26 participants, 18 male, mean age 32.8 years, SD = 8.28, 23 right-handed.

3.2 Results - Experiments 2a and 2b

Each participant’s weighted mean motion overestimation scores were entered into a $2 \times 2 \times 3$ repeated measures analysis of variance (ANOVA) with factors Action Direction (Reach vs. Withdraw), Sound (-100 ms vs. $+100$ ms) and End Position of motion termination relative to the centre of the participant’s screen on the x-axis

(Centre, Middle, Outer).

As preregistered for Experiment 2b (<https://aspredicted.org/blind.php?x=MTJ.ZNM>) our hypothesis is that participants will generally overestimate the robot’s motion, particularly during the longer sound duration, compared to the shorter sound duration. General overestimation – that participants’ perception is biased towards their prediction of the future location of the robot hand – is characterised by the intercept term in our ANOVA model, with generally positive values indicating over-estimation of motion towards the predicted next steps and negative values indicating under-estimation. The crucial hypothesis that longer consequential sounds will prompt participants to perceive the hand further into its future trajectory than shorter sounds is represented by the main effect of Sound.

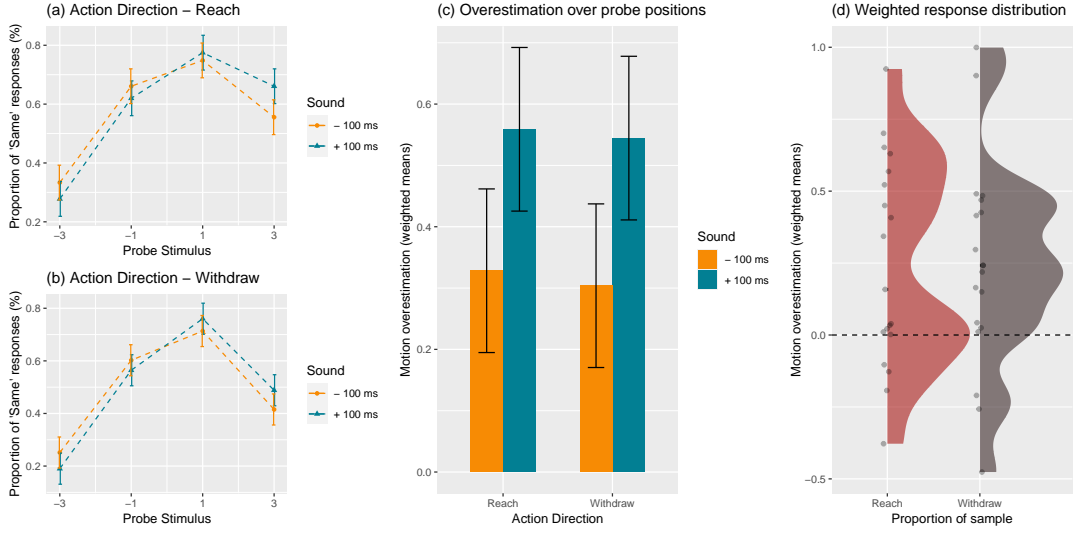
As with Experiments 1a and 1b, since we have no additional predictions, all other main effects and interactions are treated as incidental unless they meet the adjusted alpha threshold of .007 to correct for incidental findings in multi-factor ANOVAs [9].

3.2.1 Experiment 2a

Our ANOVA model revealed the predicted effect that participants’ perceptual responses were biased towards probes representing the robot’s future trajectory, as opposed to probes representing the recent past, characterised by a significant intercept ($F(1, 18) = 16.6, p < .001, n_p^2 = .480$). For both reaches (Figure 4 (a), and withdrawals (Figure 4 (b)), participants were therefore more likely to identify probes in a future location as “same” than probes in a past location, consistent with the general overestimation of perceived motion into its predicted future location (representational momentum [21, 24, 43]).

Importantly, our main prediction – that consequential sounds that terminated after the shown motion would increase this bias towards identifying future probe locations as “same” – was confirmed by the main effect of Sound ($F(1, 18) = 15.4, p < .001, n_p^2 = 0.461$), indicating larger overestimations for longer than shorter

Experiment 2a



Experiment 2b

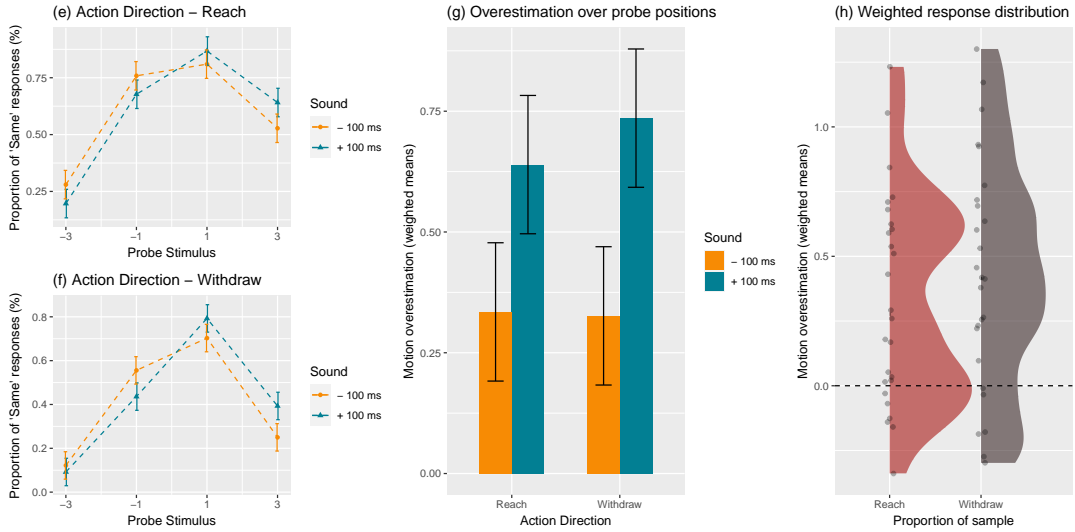


Fig. 4: For Experiment 2a, (a) and (b) show the proportion of trials in which the participant judged the probe position to be the same as the final position of the robot is plotted at each level of the probe stimulus (-3, -1, 1, 3) for Reaches and Withdraws, respectively. In (c) motion overestimation is expressed as a weighted mean. In (d), each dot represents each participant's difference in their weighted mean motion overestimation scores between the longer sound and shorter sound conditions, while the cloud presents the density of the scores for the sample. Error bars in (a), (b) and (c) represent 95% confidence intervals for the main effect of Sound. For Experiment 2b, the same is shown via (e), (f), (g) and (h)

sounds (see Figure 4 (c)).

In addition to these predicted findings, the ANOVA also revealed an interaction between Action Direction and End Position

($F(2, 36) = 8.77, p < .001, \eta_p^2 = 0.326$), which is independent of our main hypotheses. It reflects that overestimation is more pronounced for the shorter than the longer movement durations, this difference is more pronounced for withdrawals

relative to reaches.

Our ANOVA model indicated no additional main effects or interactions that met the adjusted threshold ($F < 4.99, p > 0.012$).

3.2.2 Experiment 2b

Experiment 2b successfully replicated all relevant findings of Experiment 2a. The ANOVA first indicated a significant intercept ($F(1, 25) = 62.5, p < .001, n_p^2 = 0.714$). As in Experiment 2a, participants' responses were generally biased towards probes representing the future steps of the robot motion kinematics (Figure 4 (e), (f) and (g)). Moreover, as in Experiment 2a, the ANOVA revealed the crucial main effect of Sound ($F(1, 25) = 28.7, p < .001, n_p^2 = 0.534$), confirming that longer sounds increased the overestimation of perceived motion relative to shorter sounds (reflected in Figure 4 (e), (f) and (g)).

Next to these predicted results, Experiment 2b also replicated the interaction between Action Direction and End Position, ($F(2, 50) = 6.44, p = .004, n_p^2 = 0.205$), plus a main effect of End Position ($F(2, 50) = 6.21, p = .004, n_p^2 = 0.20$). This again reflects that motion overestimation is more pronounced for the motions ending in central locations, (Centre) and weaker for motions ending in the periphery (Middle and Outer) motions, specifically for withdrawals relative to reaches.

Our ANOVA model indicated no additional main effects or interactions that met the adjusted threshold ($F < 1.65, p > 0.203$).

4 General Discussion

For productive human-robot interaction, it is crucial that human partners can represent the robot's motions effectively, so that they can plan their own behaviour in response. Here, we present a task – based on well-established representational momentum designs [14, 22] in experimental psychology – that can measure the representations human observers hold of the movements they observe. We tested whether the consequential sounds that robots produce induce illusory distortions in how observers represent the lower-level

visuospatial features of their actions and anticipate their next steps. In four experiments, we showed participants brief clips of a robot reaching forward or backward and asked them to localise the hand's last seen location after it disappeared, either with a mouse cursor (Experiment 1a and 1b) or by matching it to comparison stimuli presented directly after (Experiment 2a and 2b). We manipulated the sounds that accompanied these motions so that the sounds either terminated before or after movement offset (± 100 ms). According to multisensory cue integration frameworks [11], the sounds should be integrated into the perception of the robot's movement kinematics [71], so that identical motions appear more pronounced when accompanied by a longer sound and less so when accompanied by a shorter sound.

All experiments confirmed the predicted influences of sound on participants' perceptual judgments. In Experiment 1a and 1b, participants reported that the robot's motions extended further into space – further forward for reaches, further backward for withdrawals – when accompanied by a sound that extended beyond the motion's offset compared to a sound that terminated before the offset. This bias in visual judgments was present when participants had reliable prior knowledge about the duration of the motion they would see (i.e., the motion always terminated after 410 ms, Experiment 1a), but increased when the timing of the motion offset was unpredictable (Experiment 1b). Together, these show that visual motion and sound are dynamically integrated during robot action observation [11, 71], and induce subtle distortions in how otherwise identical visual motions are perceived. Moreover, consistent with frameworks of Bayesian integration, these influences increase when the motion offset can be less reliably estimated [33, 77].

Experiments 2a and 2b confirmed, subsequently, that the human representation of robot motion is predictive, like the perception of abstract objects [23] and biological human agents [24]. When participants judged the robot hand's disappearance point against comparison stimuli, they (mis-)identified disappearance points slightly in the future than they really were – further forward for reaches, further backward

for withdrawals – with the hand’s final location. Importantly, these predictive representations of robot motion were again affected by the sound that accompanied the motion. People identified a location even further in the future as the hand’s last seen location when the sound extended beyond the motion than when it was accompanied with a shorter sound.

Together, these findings are in line with multisensory cue integration frameworks of robot motion perception [11]. Accordingly, when viewing motion, the brain integrates all available information from the same or other modalities (here: audition), as well as prior expectations about how the motion is most likely to continue. As a consequence, people’s representation of observed motion is not veridical, but slightly biased away from what was really observed, towards the expected next steps in the motion sequence [21, 24, 43], and towards information provided by other channels. It identifies the consequential sound that accompanies robot motion – an often ignored design parameter – as a crucial component that shapes the generation of these integrated predictive percepts of robot motion.

It is remarkable that the relatively subtle manipulation of sound timing (of ± 100 ms), which remained undetected by the majority of participants even when explicitly asked about any such changes (See in the supplementary information, testing for demand effects), can induce biases of large effect size in motion judgments, indicating practical relevance. Moreover, the manipulation of sound duration affected not only the control of one’s own movements in space when guiding them to the robot’s hand, which are probed by the mouse localisation judgment task of Experiment 1a and 1b. In addition, it affected the more fundamental perceptual representation of the observed motions, which is tested by the probe judgment task of Experiment 2a and 2b, and which represents motions not where they are now, but where they will be in the next moment in time. Demonstrating such a perceptual representation is crucial because any change to how robot motions are fundamentally represented will affect all further processes that build on these representations, such as higher-level socioemotive judgments or mechanisms involved in the control

of one’s own actions in response to what the robot was seen doing [7].

The current findings go beyond prior work that has shown that the consequential sound robots produce can affect the psychometric assessment of its higher-level socioemotive characteristics, such as whether people feel “safe” when interacting with it [73], how they rate its quality, or its competence [62]. Ours is the first study to show that a robot’s consequential sound affects the perception of even low-level features of its behaviour, such as the kinematics of its movements and its expected next steps. Establishing an influence on the representation of low-level features is important because low-level visuospatial features are what ultimately informs the planning of human cooperation partners’ own actions in response within the common workspace. For example, when accepting an object from the robot, giving an object to it, or when simply navigating around it [68]. Moreover, the perception of the low-level features – such as the smoothness of a motion, its speed and its extent in space – feeds directly into higher-level judgments of more global aspects of its behaviour (e.g., competence, safety). Indeed, there is evidence that problematic emotional user responses, such as those in troughs of the uncanny valley [49], often originate from mismatches between low-level features and higher-level impressions (e.g., when a robot’s non-biological facial motions contrast with its human-like appearance, [45], or voice [47]).

While consequential sound is often neglected in robot design and interaction research, the present findings show that the sound that accompanies a robot’s actions is both a crucial issue that needs to be carefully considered in its design, and an important tool that designers and engineers have at their disposal to better integrate its behaviour into workflows spread across artificial and human operators. For applications such as in flexible production lines in which humans and robots work together and share the workspace [35, 64], the accurate representation of motion extent is of direct relevance, and consequential sounds that accompany the onset of such motions, but do not continue to their end are frequent (e.g., because of friction of moving parts, starting

up of motors). Our findings suggest that such situations are likely to induce distortions to how the robot’s behaviour is represented, and lead to problems with interactions, on both the level of more global evaluations (e.g., in terms of the robot’s predictability, jerkiness, and safety) and online action planning (e.g., where to reach for when accepting a tool from the robot).

On a theoretical and methodological level, the present findings translate fundamental psychophysical work on the integration of sound and motion into an HRI context. Prior psychophysical work outside HRI has shown that the representation of moving stimuli can be affected by the sound that accompanies them [71]. However, these studies used abstract stimuli, with limited spatial extent, and sound was manipulated between longer blocks of trials, giving room for longer-term expectations of the motions one will see and the sounds one hears across several action instances. Our research shows that sound affects motion perception in much more ecological contexts, with dynamic, naturalistic, spatially extended realistic robot parts with shading and colour, and where the sound that accompanies the motions changes dynamically with every trial, instead of longer experimental blocks. The findings therefore show that fundamental principles of human multisensory integration [11] provide an effective framework to understand – and manipulate – how robot behaviour is perceived. These frameworks assume that sensory representations, and the subjective perceptual experiences that result from them, are abstracted in probabilistic terms, in terms of likelihoods attributed to each sensory characteristic. These likelihoods are further constrained by concurrent information from other channels [71, 72], and by prior expectations about the object’s forthcoming behaviour [24], with an increased weighting of such influences when sensory information is uncertain, providing an optimally likely estimation of the relevant feature given all sensory inputs [29]. As long as the observer can infer a causal (or statistical) relationship between both cues, the perception of one should influence that of the other, as seen in both experiments [77].

While the experimental paradigm relies on video representations of robot avatars, prior

research suggests that the biases it measures feed directly into the action planning mechanisms that we use to interact with moving objects and dynamic interaction partners such as robots [51, 53]. Our experimental paradigm provides a novel tool to reveal such induced perceptual errors and to test potential remedies. Further research could compare, for example, whether different sound features – e.g., quieter compared to more energetic sounds, sounds that appear smoother or jerkier and stuttering – induce similar distortions, and whether motion perception is also affected by other visual cues, such as kinematic profiles, robot morphology, and the robot’s social cues. It has been shown for example that robots that obey biologically inspired kinematic trajectories (e.g., the two thirds power law, minimum jerk motion profiles) are easier to teleoperate /remote control [32, 69] and elicit stronger attributions of goal-directness [59]. One promising use case of the findings of this study would be teleoperation. Supplementary sound (such as that designed by Robinson and colleagues [62]) could “sonify” a teleoperated robot’s motion and act as implicit feedback for operators in circumstances where precision is paramount.

5 Conclusions

This study demonstrated that sound can be used to systematically affect how a robot’s action kinematics are visuospatially represented. A sound with a positive offset (longer sound) elicited shift in perception in the direction of motion, while negative offset (shorter sound) evoked a shift in perception against the direction of motion. The findings show that frameworks of Bayesian multisensory integration can productively be applied to robot-human-interaction research and provide an effective framework to design multimodal interactions of artificial agents (robotic and digital). We see both the robust nature of this effect and its large effect size as reason to exploit this effect in physical robotic systems to improve the perceptibility of robotic agents to humans. This study therefore contributes a foundational experimental design to test the influence of robotic design parameters on human perception, which can be built on in further studies. The Bayesian cue integration framework enables designers to create

more naturalistic interactions with designed environments and artificial agents in augmented and virtual reality. We argue that the same framework can be exploited for social robotics, and that the methodology used in this study offers a first example for deriving quantifiable relationship between modifiable design parameters and low-level human perception. These results and methodology could lead to easy-to-implement methods that aid the human representation of the teleoperated robot’s motions. This could accelerate innovation in the field of robot remote operations, ultimately leading to increased operational safety and effectiveness.

Funding. This work was supported by funding from the SGSSS/ESRC Steers Competition 2020/21 (ES/P000681/1, 2605775) to PB and MEG, and Leverhulme Trust grant RPG-2019-248 to PB.

Data Availability. The experimental files, analysis scripts and datasets generated during and/or analyzed during the current study are available from the corresponding author’s github repository:

<https://github.com/jwgcurrie/Robot-action-perception-consequential-sound>.

Statements and Declarations

Ethics. The experiments conducted for the present research were approved by the University of Aberdeen’s ethics board, and adhere to the ethical principles stated by the British Psychological Society (BPS) Code of Ethics and Conduct for psychological research with human participants. Before the experiment all individuals were informed about the basic intent of the study and gave their informed consent to participate.

Consent to publish. Participants provided informed consent regarding the publication of their anonymised data, and consented to storage of their anonymised data on an online repository.

Competing Interests of Interest. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- [1] Andersson S, Granlund A, Hedelind M, et al (2020) Exploring the Capabilities of Industrial Collaborative Robot Applications, IOS Press, pp 109 – 118. <https://doi.org/10.3233/ATDE200148>
- [2] Antoine Vandenberghe GV (2023) Predictive extrapolation of observed body movements is tuned by knowledge of the body biomechanics. *Journal of experimental psychology Human perception and performance* 49(2):188–196. <https://doi.org/https://doi.org/10.1037/xhp0001077>, URL <https://pubmed.ncbi.nlm.nih.gov/36442047/>
- [3] audacityteam (2022) Audacity. URL <https://www.audacityteam.org/>
- [4] Bach P, Schenke KC (2017) Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass* 11(7):e12312. <https://doi.org/10.1111/spc3.12312>
- [5] Bao Y, Cuijpers RH (2017) On the imitation of goal directed movements of a humanoid robot. *International Journal of Social Robotics* 9(5):691–703. <https://doi.org/10.1007/s12369-017-0417-8>
- [6] Belhassein K, Fernández-Castro V, Mayima A, et al (2022) Addressing joint action challenges in HRI: Insights from psychology and philosophy. *Acta Psychologica* 222:103476. <https://doi.org/10.1016/j.actpsy.2021.103476>
- [7] Clark A (2013) Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3):181–204. <https://doi.org/10.1017/S0140525X12000477>
- [8] Coren S, Hoenig P (1972) Effect of non-target stimuli upon length of voluntary saccades. *Perceptual and Motor Skills* 34(2):499–508. <https://doi.org/10.2466/pms.1972.34.2.499>

- [9] Cramer AOJ, van Ravenzwaaij D, Matzke D, et al (2015) Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review* 23(2):640–647. <https://doi.org/10.3758/s13423-015-0913-5>
- [10] de’Angelis N, Khan J, Marchegiani F, et al (2022) Robotic surgery in emergency setting: 2021 WSES position paper. *World Journal of Emergency Surgery* 17(1). <https://doi.org/10.1186/s13017-022-00410-6>
- [11] Ernst MO, Bühlhoff HH (2004) Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8(4):162–169. <https://doi.org/10.1016/j.tics.2004.02.002>
- [12] Fischer J, Whitney D (2014) Serial dependence in visual perception. *Nature Neuroscience* 17(5):738–743. <https://doi.org/10.1038/nn.3689>, URL <https://doi.org/10.1038/nn.3689>
- [13] Flach R, Knoblich G, Prinz W (2004) The two-thirds power law in motion perception. *Visual Cognition* 11(4):461–481. <https://doi.org/10.1080/13506280344000392>
- [14] Freyd JJ, Finke RA (1984) Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10(1):126–132. <https://doi.org/10.1037/0278-7393.10.1.126>
- [15] Gandolfo M, Downing PE (2019) Causal evidence for expression of perceptual expectations in category-selective extrastriate regions. *Current Biology* 29(15):2496–2500.e3. <https://doi.org/https://doi.org/10.1016/j.cub.2019.06.024>, URL <https://www.sciencedirect.com/science/article/pii/S0960982219307171>
- [16] Hafri A, Boger T, Firestone C (2022) Melting ice with your mind: Representational momentum for physical states. *Psychological Science* 33(5):725–735. <https://doi.org/10.1177/09567976211051744>
- [17] Hayes AE, Freyd JJ (2002) Representational momentum when attention is divided. *Visual Cognition* 9(1-2):8–27. <https://doi.org/10.1080/13506280143000296>, URL <https://doi.org/10.1080/13506280143000296>
- [18] Hogendoorn H (2020) Motion extrapolation in visual processing: Lessons from 25 years of flash-lag debate. *The Journal of Neuroscience* 40(30):5698–5705. <https://doi.org/10.1523/jneurosci.0275-20.2020>
- [19] Hogendoorn H (2020) Motion extrapolation in visual processing: Lessons from 25 years of flash-lag debate. *Journal of Neuroscience* 40(30):5698–5705. <https://doi.org/10.1523/JNEUROSCI.0275-20.2020>, URL <https://www.jneurosci.org/content/40/30/5698>, <https://www.jneurosci.org/content/40/30/5698.full.pdf>
- [20] Hogendoorn H, Burkitt AN (2019) Predictive coding with neural transmission delays: A real-time temporal alignment hypothesis. *eNeuro* 6(2):ENEURO.0412–18.2019
- [21] Hubbard TL (2005) Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review* 12(5):822–851. <https://doi.org/10.3758/bf03196775>
- [22] Hubbard TL (2010) Approaches to representational momentum: theories and models. In: *Space and Time in Perception and Action*. Cambridge University Press, p 338–365, <https://doi.org/10.1017/cbo9780511750540.020>
- [23] Hubbard TL, Courtney JR (2010) Cross-modal influences on representational momentum and representational gravity. *Perception* 39(6):851–862. <https://doi.org/10.1068/p6538>
- [24] Hudson M, Nicholson T, Ellis R, et al (2016) I see what you say: Prior knowledge of other’s goals automatically biases the perception of their actions. *Cognition* 146:245–250. <https://doi.org/10.1016/j.cognition.2015.09.021>

- [25] Hudson M, Nicholson T, Simpson WA, et al (2016) One step ahead: The perceived kinematics of others' actions are biased toward expected goals. *Journal of Experimental Psychology: General* 145(1):1–7. <https://doi.org/10.1037/xge0000126>
- [26] Hudson M, Bach P, Nicholson T (2018) You said you would! the predictability of other's behavior from their intentions determines predictive biases in action perception. *Journal of Experimental Psychology: Human Perception and Performance* 44(2):320–335. <https://doi.org/10.1037/xhp0000451>
- [27] Hudson M, McDonough KL, Edwards R, et al (2018) Perceptual teleology: expectations of action efficiency bias social perception. *Proceedings of the Royal Society B: Biological Sciences* 285(1884):20180638. <https://doi.org/10.1098/rspb.2018.0638>
- [28] Kamino W, Sabanovic S (2023) Coffee, tea, robots? the performative staging of service robots in 'robot cafes' in japan. In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, HRI '23, p 183–191, <https://doi.org/10.1145/3568162.3576967>, URL <https://doi.org/10.1145/3568162.3576967>
- [29] Kayser C, Shams L (2015) Multisensory causal inference in the brain. *PLOS Biology* 13(2):e1002075. <https://doi.org/10.1371/journal.pbio.1002075>
- [30] Kerzel D (2003) Mental extrapolation of target position is strongest with weak motion signals and motor responses. *Vision Research* 43(25):2623–2635. [https://doi.org/https://doi.org/10.1016/S0042-6989\(03\)00466-8](https://doi.org/https://doi.org/10.1016/S0042-6989(03)00466-8), URL <https://www.sciencedirect.com/science/article/pii/S0042698903004668>
- [31] Knill DC, Pouget A (2004) The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* 27(12):712–719. <https://doi.org/10.1016/j.tins.2004.10.007>
- [32] Kupferberg A, Glasauer S, Huber M, et al (2011) Biological movement increases acceptance of humanoid robots as human partners in motor interaction. *AI & SOCIETY* 26(4):339–345. <https://doi.org/10.1007/s00146-010-0314-2>
- [33] Körding KP, Beierholm U, Ma WJ, et al (2007) Causal inference in multisensory perception. *PLoS ONE* 2(9):e943. <https://doi.org/10.1371/journal.pone.0000943>
- [34] Lawrence MA (2016) ez: Easy Analysis and Visualization of Factorial Experiments. URL <https://CRAN.R-project.org/package=ez>, r package version 4.4-0
- [35] Liu C, Tomizuka M (2016) Algorithmic safety measures for intelligent industrial co-robots. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, <https://doi.org/10.1109/icra.2016.7487476>
- [36] Liu H, Wang L (2018) Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics* 68:355–367. <https://doi.org/10.1016/j.ergon.2017.02.004>
- [37] Lum PS, Burgar CG, Shor PC, et al (2002) Robot-assisted movement training compared with conventional therapy techniques for the rehabilitation of upper-limb motor function after stroke. *Archives of Physical Medicine and Rehabilitation* 83(7):952–959. <https://doi.org/10.1053/apmr.2001.33101>
- [38] MacDorman KF, Chattopadhyay D (2016) Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146:190–205. <https://doi.org/10.1016/j.cognition.2015.09.019>
- [39] Manassi M, Liberman A, Kosovicheva A, et al (2018) Serial dependence in position occurs at the time of perception. *Psychonomic Bulletin & Review* 25(6):2245–2253. <https://doi.org/10.3758/s13423-018-1454-5>
- [40] Marge M, Espy-Wilson C, Ward NG, et al (2022) Spoken language interaction with

- robots: Recommendations for future research. *Computer Speech & Language* 71:101255. <https://doi.org/10.1016/j.csl.2021.101255>
- [41] MathWorks (2022) Matlab2022a. URL <https://uk.mathworks.com/products/matlab.html>
- [42] Maurice P, Huber ME, Hogan N, et al (2018) Velocity-curvature patterns limit human-robot physical interaction. *IEEE Robotics and Automation Letters* 3(1):249–256. <https://doi.org/10.1109/lra.2017.2737048>
- [43] McDonough KL, Hudson M, Bach P (2019) Cues to intention bias action perception toward the most efficient trajectory. *Scientific Reports* 9(1). <https://doi.org/10.1038/s41598-019-42204-y>
- [44] McDonough KL, Costantini M, Hudson M, et al (2020) Affordance matching predictively shapes the perceptual representation of others’ ongoing actions. *Journal of Experimental Psychology: Human Perception and Performance* 46(8):847–859. <https://doi.org/10.1037/xhp0000745>
- [45] Meding K, Bruijns SA, Schölkopf B, et al (2020) Phenomenal causality and sensory realism. *i-Perception* 11(3):204166952092703. <https://doi.org/10.1177/2041669520927038>
- [46] millisecond (2022) Inquisit6. URL <https://www.millisecond.com/>
- [47] Mitchell WJ, Kevin A Szerszen S, Lu AS, et al (2011) A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2(1):10–12. <https://doi.org/10.1068/i0415>, URL <https://doi.org/10.1068/i0415>, pMID: 23145223, <https://doi.org/10.1068/i0415>
- [48] Moon I, Joung S, Kum Y (2002) Safe and reliable intelligent wheelchair robot with human robot interaction. In: *Proceedings 2002 IEEE International Conference on Robotics and Automation* (Cat. No.02CH37292). IEEE, <https://doi.org/10.1109/robot.2002.1014267>
- [49] Mori M (2017) The uncanny valley: The original essay by masahiro mori. *IEEE Robots &*
- [50] Müsseler J, der Heijden AHCv (2004) Two spatial maps for perceived visual space: Evidence from relative mislocalizations. *Visual Cognition* 11(2-3):235–254. <https://doi.org/10.1080/13506280344000338>, URL <https://doi.org/10.1080/13506280344000338>, <https://doi.org/10.1080/13506280344000338>
- [51] Müsseler J, Stork S, Kerzel D (2008) Localizing the onset of moving stimuli by pointing or relative judgment: Variations in the size of the fröhlich effect. *Vision Research* 48(4):611–617. <https://doi.org/10.1016/j.visres.2007.11.024>
- [52] Nanavati A, Alves-Oliveira P, Schrenk T, et al (2023) Design principles for robot-assisted feeding in social contexts. In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, HRI ’23, p 24–33, <https://doi.org/10.1145/3568162.3576988>, URL <https://doi.org/10.1145/3568162.3576988>
- [53] Nijhawan R (1994) Motion extrapolation in catching. *Nature* 370:256–257
- [54] Nijhawan R (2002) Neural delays, visual motion and the flash-lag effect. *Trends in Cognitive Sciences* 6(9):387–393. [https://doi.org/https://doi.org/10.1016/S1364-6613\(02\)01963-0](https://doi.org/https://doi.org/10.1016/S1364-6613(02)01963-0), URL <https://www.sciencedirect.com/science/article/pii/S1364661302019630>
- [55] Pandey AK, Gelin R (2018) A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine* 25(3):40–48. <https://doi.org/10.1109/MRA.2018.2833157>
- [56] Pascucci D, Plomp G (2021) Author correction: Serial dependence and representational momentum in single-trial perceptual decisions. *Scientific Reports* 11(1). <https://doi.org/10.1038/s41598-021-96023-1>

- [57] Pitcher D, Ungerleider LG (2021) Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences* 25(2):100–110. <https://doi.org/https://doi.org/10.1016/j.tics.2020.11.006>, URL <https://www.sciencedirect.com/science/article/pii/S1364661320302783>
- [58] R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>
- [59] Richardson MJ, Flash T (2000) On the emulation of natural movements by humanoid robots. In: *IEEE-RAS International Conference on Humanoids Robots*, Citeseer
- [60] Riek LD, Rabinowitch TC, Bremner P, et al (2010) Cooperative gestures: Effective signaling for humanoid robots. In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, <https://doi.org/10.1109/hri.2010.5453266>
- [61] Ritzberger N, Emenshuber S, Schubert G, et al (2017) Mobile robot for exploration and manipulation inside a nuclear power plant. In: *DAAAM Proceedings*. DAAAM International Vienna, p 0093–100, <https://doi.org/10.2507/28th.daaam.proceedings.012>
- [62] Robinson FA, Velonaki M, Bown O (2021) Smooth operator. In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, <https://doi.org/10.1145/3434073.3444658>
- [63] Sacheli LM, Arcangeli E, Carioti D, et al (2021) Taking apart what brings us together: The role of action prediction, perspective-taking, and theory of mind in joint action. *Quarterly Journal of Experimental Psychology* 75(7):1228–1243. <https://doi.org/10.1177/17470218211050198>
- [64] Sauppé A, Mutlu B (2015) The social impact of a robot co-worker in industrial settings. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, <https://doi.org/10.1145/2702123.2702181>
- [65] Saygin AP, Chaminade T, Ishiguro H, et al (2011) The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience* 7(4):413–422. <https://doi.org/10.1093/scan/nsr025>
- [66] Schadenberg BR, Reidsma D, Heylen DKJ, et al (2021) “i see what you did there”. *ACM Transactions on Human-Robot Interaction* 10(3):1–28. <https://doi.org/10.1145/3461534>
- [67] Sebanz N, Knoblich G (2009) Prediction in joint action: What, when, and where. *Topics in Cognitive Science* 1(2):353–367. <https://doi.org/10.1111/j.1756-8765.2009.01024.x>
- [68] SEBANZ N, BEKKERING H, KNOBLICH G (2006) Joint action: bodies and minds moving together. *Trends in Cognitive Sciences* 10(2):70–76. <https://doi.org/10.1016/j.tics.2005.12.009>
- [69] Shafiei SB, Cavuoto L, Guru KA (2017) Motor skill evaluation during robot-assisted surgery. In: *Volume 5A: 41st Mechanisms and Robotics Conference*. American Society of Mechanical Engineers, <https://doi.org/10.1115/detc2017-67607>
- [70] Tennent H, Moore D, Jung M, et al (2017) Good vibrations: How consequential sounds affect perception of robotic arms. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, <https://doi.org/10.1109/roman.2017.8172414>
- [71] Teramoto W, Hidaka S, Gyoba J, et al (2010) Auditory temporal cues can modulate visual representational momentum. *Attention, Perception, & Psychophysics* 72(8):2215–2226. <https://doi.org/10.3758/bf03196696>
- [72] Teramoto W, Manaka Y, Hidaka S, et al (2010) Visual motion perception induced by sounds in vertical plane. *Neuroscience Letters* 479(3):221–225. <https://doi.org/10.1016/j.neulet.2010.05.065>

- [73] Trovato G, Do M, Terlemez O, et al (2016) Is hugging a robot weird? investigating the influence of robot appearance on users' perception of hugging. In: 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). IEEE, <https://doi.org/10.1109/humanoids.2016.7803295>
- [74] Weiss Y, Simoncelli EP, Adelson EH (2002) Motion illusions as optimal percepts. *Nature Neuroscience* 5(6):598–604. <https://doi.org/10.1038/nn0602-858>
- [75] Wykowska A, Kajopoulos J, Obando-Leitón M, et al (2015) Humans are well tuned to detecting agents among non-agents: Examining the sensitivity of human perception to behavioral characteristics of intentional systems. *International Journal of Social Robotics* 7(5):767–781. <https://doi.org/10.1007/s12369-015-0299-6>
- [76] Yang S, Bill J, Drugowitsch J, et al (2021) Human visual motion perception shows hallmarks of bayesian structural inference. *Scientific Reports* 11(1). <https://doi.org/10.1038/s41598-021-82175-7>
- [77] Yon D, Frith CD (2021) Precision and the bayesian brain. *Curr Biol* 31(17):R1026–R1032
- [78] Zhanatkyzy A, Telisheva Z, Amirova A, et al (2023) Multi-purposeful activities for robot-assisted autism therapy: What works best for children's social outcomes? In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, HRI '23, p 34–43, <https://doi.org/10.1145/3568162.3576963>, URL <https://doi.org/10.1145/3568162.3576963>