An Analysis of Deceptive Robot Motion

Anca Dragan, Rachel Holladay, and Siddhartha Srinivasa
The Robotics Institute, Carnegie Mellon University

***Abstract*—Much robotics research explores how robots can clearly communicate true information. Here, we focus on the counterpart: communicating false information, or hiding information altogether – in one word, deception. Robot deception is useful in conveying intentionality, and in making games against the robot more engaging. We study robot deception in goal-directed motion, in which the robot is concealing its actual goal. We present an analysis of deceptive motion, starting with how humans would deceive, moving to a mathematical model that enables the robot to autonomously generate deceptive motion, and ending with a study on the implications of deceptive motion for human-robot interactions.**

I. Introduction

Much robotics research explores how robots can communicate effectively, via speech [9, 19, 33], gesture [5, 24, 25, 37], or motion [1, 3, 13, 18, 22, 30]. But effective communication, which clearly conveys truthful information, has a natural counterpart: effective *deception*, which clearly conveys false information, or hides information altogether.

Robotic deception has obvious applications in the military [10], but its uses go far beyond. At its core, deception conveys *intentionality* [31], and that the robot has a *theory of mind* for the deceived [4] which it can use to manipulate their beliefs. It makes interactions with robots more engaging, particularly during game scenarios [28, 31, 32].

Among numerous channels for deception, we focus on deception via *motion*. Deceptive motion is an integral part of being an opponent in most sports, like squash [16], soccer [29], or rugby [21]. It can also find uses outside of competitions, such as tricking patients into exerting more force during physical therapy [6]. Furthermore, a robot that can generate deceptive motion also has the ability to quantify an accidental leakage of deception and therefore avoid deceiving accidentally.

We study deception in *goal-directed* motion, where a robot is moving towards one of a few candidate goals — we refer to this one as the robot's *actual* goal. Fig.1 shows an example: the robot is reaching for one of two bottles on the table. In this context, we introduce the following definition:

Definition: *Deceptive motion is motion that tricks the observer into believing that the robot is* ***not*** *moving towards its actual goal.*

We present an analysis of deceptive goal-directed robot motion through a series of five user studies, from how humans would deceive, to how a robot can plan deceptive motion, to what implications this has for human-robot interactions. We make the following contributions:
1. Human Deception: We begin by studying what deception strategies people employ when creating deceptive motion for a robot (Sec. II).

We focus on a simple, 2D robot character, whose only channel of expression is its motion. We collect demonstrations of deceptive motion from novice users, as well as from a senior animation designer.

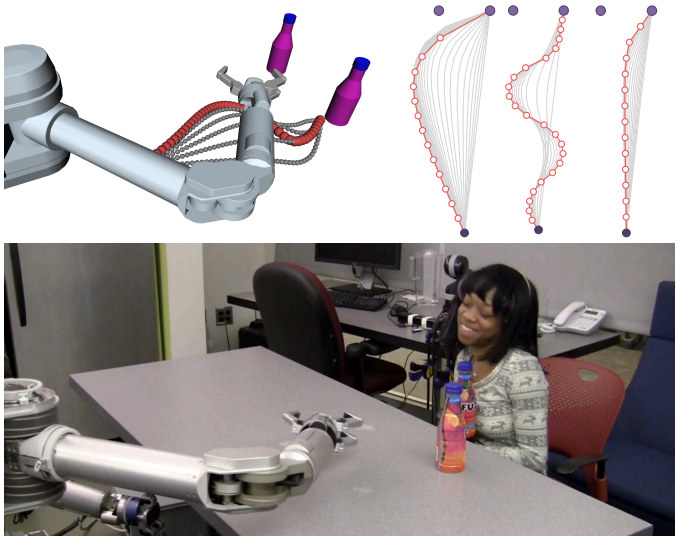

Fig. 1. Top: Deceptive motions produced by trajectory optimization. The trajectories on the right correspond to different strategies that humans adopt. Bottom: A user's reaction when she first realizes the robot deceived her about which bottle it was going to grasp.

We then cluster the demonstrations to reveal common strategies, and relate the emerging strategies to the theory of deceptive behavior in humans [36]. We find both strategies meant to "show the false" (e.g., convey a different goal), as well as strategies meant to "hide the truth" (e.g., keep the goal ambiguous until the end).
2. Mathematical Model: Next, we introduce a mathematical model for autonomously generating deceptive motion (Sec. III), and show how different parameters lead to the different user strategies revealed in the study.

Our approach is complementary to existing methods for autonomous deception, which usually lie at the symbolic level, and are inspired by either game theory [34, 35] or biology [15, 26].

Fig.1(top) shows three examples generated by our model: a trajectory that conveys the wrong goal (along with its higher-dimensional counterpart on the left), one that switches between conveying either goal, and one that keeps the goal as ambiguous as possible.
3. Evaluation: We test whether novice users are actually deceived by the robot, when executing the trajectories from the model, users, and animator (Sec. IV). We find that all motions are significantly more deceptive than a baseline, and that the model performs almost equivalently to the animator trajectory.
4. Generalization: We show how our model generalizes to higher-DOF robots – manipulator arms (Sec. V). We verify its ability to deceive in a quantitative user study, and we compare the output trajectories qualitatively to the arm motions produced by humans when asked to

deceive.

5. Implications for HRI: Our work investigates deceptive motion and proposes a model that enables robots to autonomously generate it. Our final study explores the implications of the robot's deception through its motion to human-robot interaction (Sec. VI, and Fig.1(bottom)). We focus on whether users realize that the robot is intentionally deceiving them, and, if so, whether that impacts their perceptions of the robot, be it positively or negatively.

Overall, we see that deceptive motion can improve perceptions about the robot, but can also harm trust, especially when interpreted as intentional. An additional pitfall is that some users rationalize the robot's behavior, not believing that a robot would be capable of intentional deception.

On the one hand, our work echoes concerns raised in prior work about the effects of deception [20]. On the other hand, deception is human nature. So much so that even the Turing Test is about deception — we only consider computers to have achieved human-like intelligence once they can deceive us that they are human. We are excited to contribute to a better understanding of this behavior.

II. Study 1: How Do Humans Deceive?

Humans are experts in deception. Thus, we begin our analysis of deceptive robot motion by looking to humans, and how they think a robot should deceive through its motion. We investigated this for the simple case of a 2D robot character. Without eyes and facial expression, all this character has available to deceive is its motion. We asked novice users, as well as an animator – an expert in designing expressive motion for non-humans – to provide demonstrations of deceptive motion in various scenarios.

A. Experimental Setup

Scenarios. We designed our scenarios by manipulating several factors. We start from a canonical scenario: a starting position and two candidate equidistant goals, placed close to each other to make the direct motion ambiguous, like in Fig.2a.

From this scenario, in order to test how the trajectories change when different aspects of the environment change, we manipulate: (1) *goal side*, by changing which of the goals is the actual goal; (2) *scale*, by scaling the entire scene; (3) *distance*, by translating both goals vertically; (4) *ambiguity*, by translating both goals horizontally; (5) *goal symmetry*, by shifting one goal up and one down.

Additionally, we added *multiple goals* scenarios with three candidate goals instead of two, like in Fig.3d, and scenarios in which one goal is in front of the other, for a total of 11 scenarios.

Procedure. We developed a graphical interface for providing demonstrations by placing waypoints along the trajectory. For each scenario, we first asked users to demonstrate a typical (predictable) trajectory to a goal (how they would normally expect the robot to move), in order to check that all users are working with the same underlying model of the robot motion. All users drew a straight line from start to goal (we use this in our model from Sec. III).

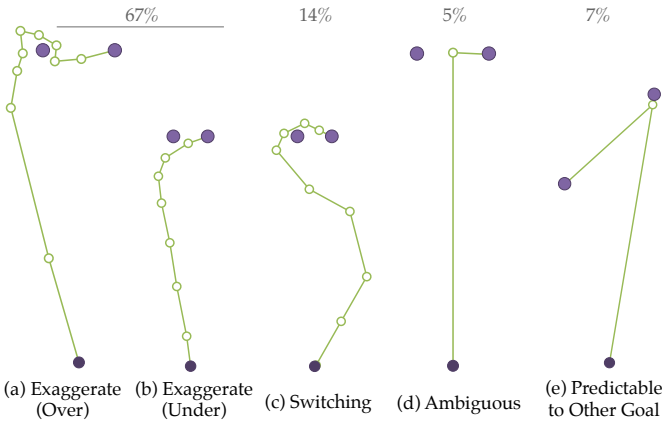

Fig. 2. User strategies for deception. The typical strategy exaggerates in the other direction and avoids the obstacle by going over it. A less common strategy of going under the obstacle closely matches the result of the model we use in Sec. III, shown in Fig.6 (red).

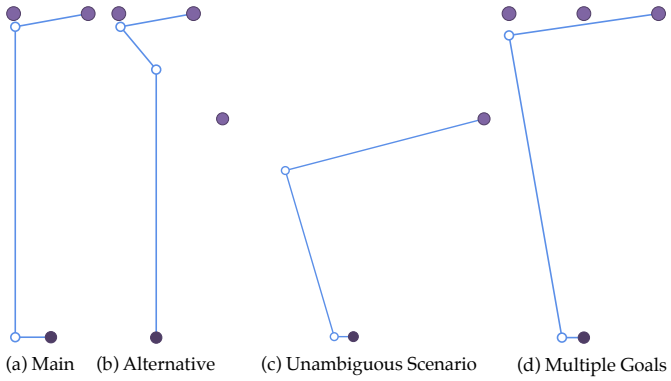

Fig. 3. Animator strategies.

Next, the users demonstrated the deceptive trajectory and explained their strategy, including how they would time the motion.

For each user, we randomized the order of the scenarios after the canonical one to avoid ordering biases. We kept the more complex multi-goal scenarios for the end.

Participants. We recruited 6 participants from the local community (4 male, 2 female, aged 19-67, with various educational backgrounds), along with a senior animation designer who we treat as an expert.

B. Analysis

Main Scenarios. We started from the user comments, and identified 4 emerging strategies (one with 2 variations), shown in Fig.2. We then classified each trajectory as employing one of the strategies, or doing something different (e.g. moving "as if the robot is broken"). We tested agreement between two coders with Cohen's κ ($\kappa = .8$, $p < .0001$).

By far, the most common strategy (67% of the cases) was to *exaggerate* the motion towards another candidate goal in order to convey the intention to reach that goal to the observer. This type of behavior closely resembles *decoying* in human deception theory [36]: it is a way of portraying false information (that the robot has a differ-

ent goal from its actual goal) by offering a misleading alternate option.

Among trajectories that follow this strategy, most (71%) avoid the other goal by going over the top, like in Fig.2a. Often, the trajectories circle this goal first (a behavior some of the users described as simulating "hovering"), and then move on to the actual goal. The rest use the more efficient (shorter) strategy of avoiding the other goal by moving under it, as in Fig.2b.

The other three strategies were drastically less common. In 14% of the cases, the users were *switching* between conveying the actual goal and conveying a different one, as in Fig.2c. This most closely resembles deception by *dazzling* [36], which is hiding the true by being confusing.

Approximately 5% of trajectories were *ambiguous* (Fig.2d), trying to conceal which goal is the actual one for as long as possible. This "hiding the real" behavior is known as *masking* in human deception literature [36], whereby "all distinctive characteristics" of the motion are concealed.

Finally, another 7% of the trajectories simply moved to the other goal, without exaggerating, and then moved to the actual goal, as in Fig.2e — this can be thought of as a variation on decoying.

Multiple Goals. When there are multiple goals in the scene, our main question was whether users would pick a particular other goal to convey, or whether they will be ambiguous in the general direction of the other goals. However, some users surprised us with an unexpected strategy: conveying one of the other goals first, then another, and only then moving towards the actual goal. Aside from this strategy, we found similar patterns as in the two goal case, predominantly exaggeration and switching.

Animator Demonstrations. Fig.3 shows the animator strategies for a few of the scenarios. For the canonical scenario, the robot first moves horizontally to align itself with the other goal in order to clearly indicate its (deceptive) selection, then goes towards it, and then, when it has almost reached it, moves towards the actual goal (Fig.3a).

The animator also proposed an alternative (Fig.3b), which is ambiguous for the majority of the trajectory, then switches to the wrong goal, then optionally oscillates between the two (conveying that the robot is exploring different options), and only then moves to the correct one. Although this trajectory is rich in expression, he deems the first one more deceptive because the observer will believe in the wrong goal for longer and with higher confidence.

Changes Between Scenarios. The users were surprisingly inconsistent with their strategies between different scenarios, but their comments reflect that they took the opportunity to explore "something new", and not that they thought that these different situations require different strategies.

With the animator, the strategy stays the same with scale, distance, goal side and symmetry. With less ambiguous scenarios, like in Fig.3c, the trajectory does not go as far in the direction of the other goal: the animator considered it enough to convince the observer that the robot is targeting the other goal.

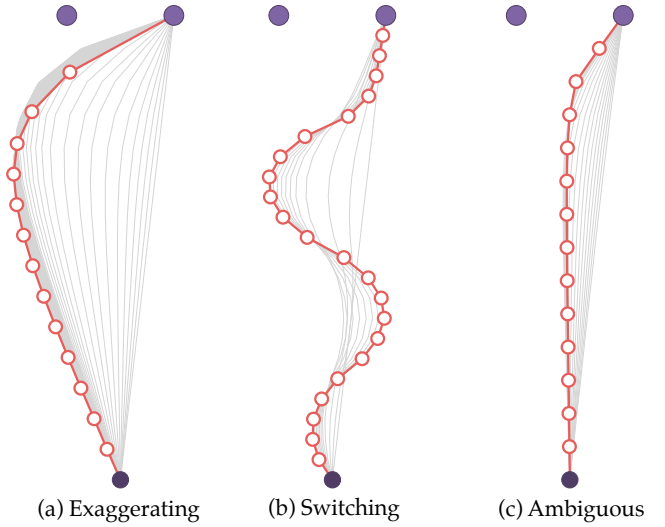

Fig. 4. Strategies replicated by the model: the typical exaggeration towards another goal, as well as the switching and ambiguous trajectories. The trajectories in gray show the optimization trace, starting from the predictable trajectory.

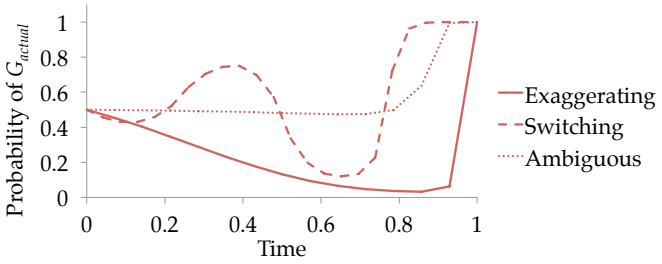

Fig. 5. The probability of the actual goal along each model trajectory.

III. A Mathematical Model for Deception

The previous section analyzed the different strategies that humans would employ to enable a robot to deceive. Here, we introduce a mathematical model for deceptive motion that (1) enables a robot to autonomously generate deceptive motion, and (2) gives us more insight into the human strategies.

Deceptive Motion as Trajectory Optimization. Our model for deceptive motion is about enabling the robot to take the observer's perspective, and compute what goal they would infer from the motion. Then, the robot can choose a motion that prevents the observer from inferring the correct goal.

Based on prior work [11], we model the observer as expecting the robot to move optimally, optimizing some efficiency cost functional $C : \Xi \rightarrow \mathbb{R}^+$ defined over the space of trajectories Ξ. We then approximate the probability of a candidate goal G being inferred from an ongoing motion $\xi_{S \rightarrow Q}$, from the start S to the current robot configuration Q, as in [12]:

$$P(G|\xi_{S\rightarrow Q}) = \frac{1}{Z}\frac{\exp\left(-C[\xi_{S\rightarrow Q}] - V_G(Q)\right)}{\exp\left(-V_G(S)\right)}P(G) \qquad (1)$$

with Z a normalizer across the set of candidate goals $\mathcal{G}$ and $V_G(q) = \min_{\xi \in \Xi_{q \to G}} C[\xi]$. This computes how costly reaching the goal is through the ongoing trajectory relative to the optimal way, and matches teleological reasoning in action interpretation theory [7, 11].

Given this, the robot can be most deceptive along the trajectory by minimizing the probability that the actual goal, G_{actual}, will be inferred:

$$\min_{\xi} \int P(G_{actual}|\xi_{S \to \xi(t)})dt \tag{2}$$

Solution. We solve this trajectory optimization problem using functional gradient descent, following [13, 38]: we minimize the first order approximation of the objective plus a regularization term that keeps the trajectory smooth.

We avoid collisions with obstacles by adding a constraint that penalizes small distances between the robot and the obstacle. We use the obstacle term from the CHOMP trajectory optimizer [38], and follow the derivation from [13] for trust region constraints to keep this term under a desired threshold.

For the cost C, we use a common choice in trajectory optimization – the integral over squared velocities [38]. This has been shown to match what users expect for a 2DOF robot [11], and to have a high degree of predictability for a 7DOF robot arm as well [14].

Strategies. Using this formalism, we can model the different user strategies from the previous sections.

The *typical* user strategy is about selecting another goal, G_{decoy}, and conveying that through the motion. In our model, this translates to maximizing the probability of that goal:

$$\xi_{exaggerate} = \arg\max_{\xi} \int P(G_{decoy}|\xi_{S \to \xi(t)})dt \tag{3}$$

When there are only two candidate goals, this is equivalent to (2).

Solving this optimization problem leads to the trajectory in Fig.4a, which qualitatively replicates the strategy in Fig.2b of exaggerating the motion towards the other goal.

The *predictable-to-other-goal* strategy in Fig.2e is similar, but instead of exaggerating, the robot moves predictably. However, prior work in conveying goals [11] has shown exaggeration to be more effective.

The animator's main demonstration (Fig.3a) follows an idea similar to exaggeration, except that conveying the goal is done through alignment – a strategy outside of the realm that our model can produce. However, in Sec. IV, we show that the model and animator trajectories perform similarly in practice.

The *switching* user trajectory (Fig.2c) alternates between the goals. If $\sigma : [0,1] \to \mathcal{G}$ is a function mapping time to which goal to convey at that time, then the switching trajectory translates in our model to maximizing the probability of goal $\sigma(t)$ at every time point:

$$\xi_{switching} = \arg\max_{\xi} \int P(\sigma(t)|\xi_{S \to \xi(t)})dt \tag{4}$$

Unlike other strategies, this one depends on the choice of σ. Optimizing for a default choice of σ (a piece-wise function alternating between G_{other} and G_{actual}, $\sigma(t) = G_{other}$ for $t \in [0,.25) \cup [.5,.75)$ and $\sigma(t) = G_{actual}$ for $t \in [.25,.5) \cup [.75,1]$) leads to the trajectory from Fig.4b, which alternates between conveying the goal on the right and the one on the left.

The *ambiguous* user trajectory (Fig.2d) keeps both goals as equally likely as possible along the way, which translates to minimizing the absolute difference between the probability of the top two goals:

$$\xi_{ambiguous} = \arg\min_{\xi} \int |P(G_{actual}|\xi_{S \to \xi(t)}) - P(G_{other}|\xi_{S \to \xi(t)}))|dt \tag{5}$$

Fig.4c is the outcome of this optimization: it keeps both goals just as likely until the end, when it commits to one. An alternate way of reaching such a strategy is to maximize the *entropy* of the probability distribution over all goals in the scene.

Using this model, we see that different strategies can be thought of as optimizing different objectives, which gives us insight into why exaggeration was so much more popular: *it is the most effective at reducing the probability of the actual goal being inferred along the trajectory*. Fig.5 plots the $P(G_{actual})$ along the way for each strategy: the lower this is, the more deceptive the strategy. While the ambiguous strategy keeps the probability distribution as close to 50-50 as possible, and the switching strategy conveys the actual goal for parts of the trajectory, the exaggerate (or decoy) strategy biases the distribution toward the other goal as much as possible for the entire trajectory duration: the observer will not only be wrong, but will be *confidently* wrong.

IV. Study 2: Are Users Really Deceived?

In this section, we compare the mathematical model and the user and animator demonstrations in terms of how deceptive they actually are (how low the probability assigned to the actual goal is) as measured with novice users. Of the three strategies in Fig.4, we use the exaggeration strategy for our comparison for two reasons: (1) it is by far the most commonly adopted strategy (69% the user demonstrations, and the main strategy for the animator); (2) it is the most deceptive – both mathematically (see Fig.5) and according to the expert animator (see Sec. II-B); the other two strategies, although interesting, are optimizing different objectives – comparing trajectories within these strategies would require fundamentally different metrics, for ambiguity and switching/confusion.

A. Experimental Setup

Hypotheses.

H1. *All 3 deceptive trajectories (the users', the animator's, and the model's) are significantly more deceptive than the predictable baseline.*

H2. *All 3 deceptive trajectories are equivalently deceptive.*

Manipulated Factors. We manipulated two factors: the type of *trajectory* used (with 4 levels), and the *time point* at which the trajectory is evaluated (with 3 levels), leading to a total of 12 conditions.

We used the typical user trajectory from Fig.2a, the main animator trajectory from Fig.3a, the output of the

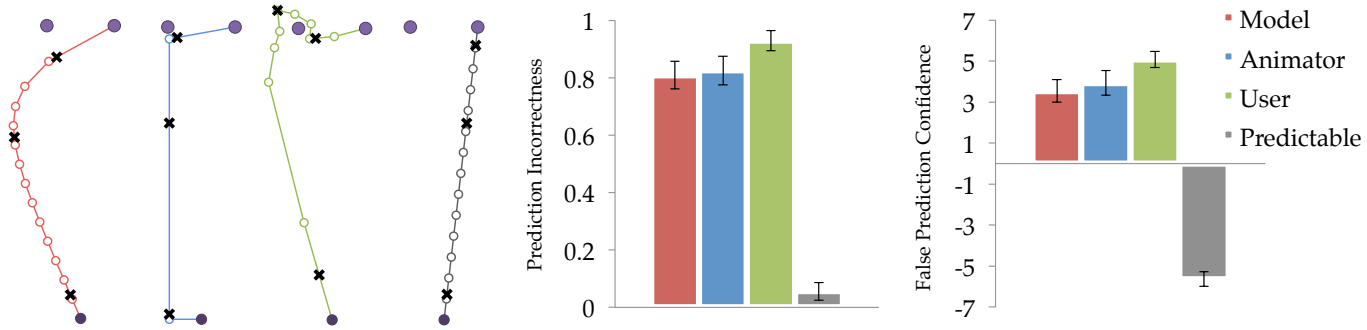

Fig. 6. The four trajectories: model, animator, user, and the predictable baseline, along with the comparison from our user study.

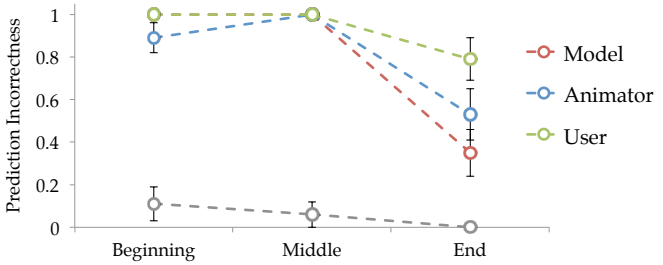

Fig. 7. A comparison of the four trajectories in terms of how deceptive they are across the three time points.

model from Fig.4a, and the predictable (straight line) motion as a baseline. Because the situation is somewhat ambiguous, the predictable trajectory does not give away the actual goal immediately.

We timed the trajectories such that they all take the same total time to execute, and followed their designers' instructions for which parts should be faster or slower. For the model trajectory, we treated each waypoint as equally spaced in time.

We selected three critical time points for evaluating the trajectories that best capture their differences: one close to the beginning of the trajectory (right after the robot executing the animator trajectory has finished aligning with the other goal), one close to the end, after all trajectories have started moving in the direction of the actual goal, and one in the mid-part, when the robot executing the user trajectory has started hovering around the other goal. We mark these points in Fig.6(left), which shows the four trajectories side by side.

Dependent Measures. We measured how deceptive the trajectories are by measuring which goal the users believe the robot is going toward as the trajectory is unfolding: the less correct the users are, the more deceptive the motion.

For each trajectory and time point, we generated a video of the robot executing the trajectory up to that time point. We measured *incorrectness* and *confidence*. We asked the users to watch the video, predict which goal the robot is going towards, and rate their confidence in the prediction on a 7 point Likert scale. We treat the confidence as negative for correct predictions (meaning the trajectory failed to deceive).

Participants. We decided on a between-subjects design, where each participant would only see one trajectory snippet, in order to avoid biases arising from having seen a different condition before.

We recruited a total of 240 users (20 per condition) on Amazon's Mechanical Turk, and eliminated users who failed to answer a control question correctly, leading to 234 users (166 male, 68 female, aged 18-60).

B. Analysis

A factorial ANOVA on *incorrectness* (considered to be robust to dichotomous data[8]) revealed significant main effects for both *trajectory* ($F(3,222) = 47.78$, $p < .0001$) and *time point* ($F(2,222) = 39.87$, $p < .0001$), as well as a significant interaction effect ($F(6,222) = 5,.5$, $p < .0001$).

The post-hoc analysis with Tukey HSD revealed two findings. (1) The predictable trajectory was significantly less deceptive than all other trajectories for all times. The one exception was the last time point of the model trajectory, which revealed the correct goal in 65% of the cases. (2) The beginning and middle time points for all three strategies were significantly more deceptive than their last time point (aside from the last time point of the user trajectory).

Fig.7 echoes these findings: it plots the mean *incorrectness* for all trajectories across the time points. The predictable trajectory deceives very few users in the beginning, and makes the actual goal more clear as time progresses. In line with **H1**, a Tukey HSD that marginalizes over time shows that the predictable trajectory is significantly less deceptive than the rest, with $p < .0001$ for all three contrasts.

Comparing the three strategies, we see that all three perform very well in the middle time point: this is expected, as by that point the robot would have been making steady progress towards the other goal. In the beginning of the trajectory, the model and the user trajectories are just as convincingly deceiving, but users actually manage to interpret the animator's trajectory as going towards the correct goal, justifying that "it seemed that the robot was veering back to the right".

The bigger differences come towards the end. The model trajectory, being smoother, gives away the actual goal sooner than the animator. Users recognize in their comments that "at the last second it turned towards the right". Surprisingly, the user "hovering" strategy worked very well, delaying the time when users catch on to the actual goal, and making it much more effective than the animator's strategy. Users actually used the term "hover" to describe the behavior, much like the designer of the trajectory himself.

Therefore, w.r.t. **H2**, the animator and user trajectories are not equivalent. However, there is a very small differ-

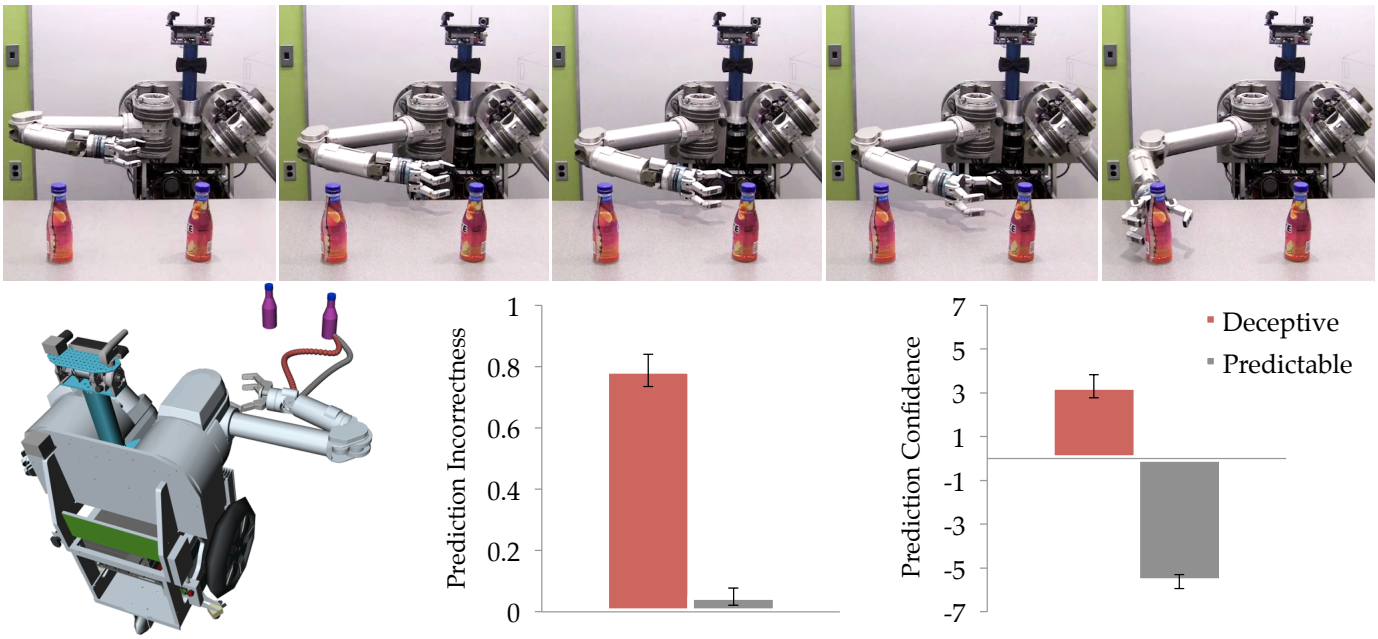

Fig. 8. Top: The deceptive trajectory planned by the model. Bottom: a comparison between this trajectory and the predictable baseline.

ence between the model and the animator trajectories, and a TOST equivalence test deems them as marginally equivalent for a practical difference threshold of 0.1 ($p = .07$).

The *confidence* metric echoes these results as well, and Fig.6 plots both. A factorial ANOVA for this measure yields analogous findings.

Overall, we see that the model (which the robot can use to autonomously generate trajectories) performs almost equivalently to the expert demonstration from the animator, and that creativity paid off for the user's "hover" strategy.

V. Generalization to Arm Motion

The previous section revealed that the mathematical model from Sec. III performs well in practice. But how well does it generalize beyond a simple 2D robot character?

In this section, we put this to the test by applying the model to the 7DOF right arm of a bi-manual mobile manipulator. Fig.8 (top) shows the resulting deceptive trajectory, along with a comparison between its end effector trace and that of the predictable trajectory (bottom left).

Both trajectories are planned s.t. they minimize cost and avoid collisions, as explained in Sec. III. The difference is in the cost functional: the predictable trajectory minimizes C (Sec. III), while the deceptive one minimizes the cost from (2). The planning time for either trajectory is based on CHOMP, and remains under a second.

Fig.1 shows the optimization trace transforming the predictable into the deceptive trajectory. After a few iterations, the trajectory shape starts bending to make progress in the objective, but remains on the constraint manifold imposed by the obstacle avoidance term.

A. Study 3: Robot Trajectory Evaluation

To evaluate whether this trajectory is really deceptive, we repeat our evaluation from Sec. IV, now with the physical robot.

Manipulated Factors and Dependent Measures. We again manipulate *trajectory* and *time-point*, this time with only two levels for the trajectory factor: the deceptive and predictable trajectories from Fig.8. This results in 6 conditions. We use the same dependent measures as before.

Participants. For this study, we recruited 120 participants (20 per condition; 80 male, 40 female, aged 19-60) on Amazon's Mechanical Turk.

Hypothesis. *The model deceptive trajectory is more deceptive than the predictable baseline.*

Analysis. In line with our hypothesis, a factorial ANOVA for *correctness* did reveal a significant main effect for *trajectory* ($F(1,117) = 150.81$, $p < .0001$). No other effects were significant. Fig.8 plots the results.

The users who were deceived relied on the principle of rational action [17], commenting that the robot's initial motion towards the left "seemed like an inefficient motion if the robot were reaching for the other bottle".

When the robot's trajectory starts moving towards the other bottle, the users find a way to rationalize it: "I think that jerking to my left was to adjust it arm to move right.", or "It looks as if the robot is going for the bottle on my right and just trying to get the correct angle and hand opening".

As for the features of the motion that people used to make their decision, the direction of the motion and the proximity to the target were by far the most prevalent, though one user quoted hand orientation as a feature as well.

Not all users were deceived, especially at the end. A few users guessed correctly from the very beginning, making (false) arguments about the robot's kinematics, e.g. "he moved the arm forward enough so that if he swung it round he could reach the bottle".

Overall, our test suggests that the model from Sec. III can generalize to higher-dimensional spaces. Next,

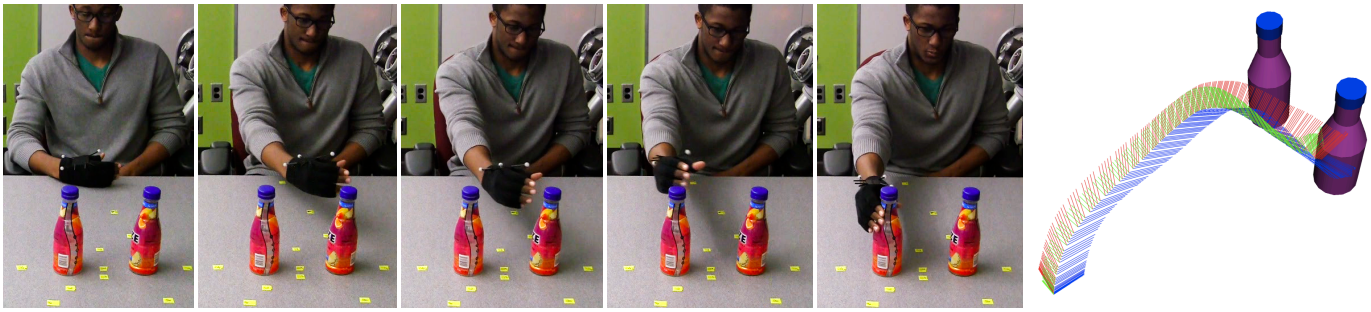

Fig. 9. A deceptive trajectory for a human arm from one of our participants, qualitatively similar to the robot trajectory from Fig.8.

we run a further user study which indicates that this was no coincidence: when we ask humans to produce deceptive motion with their own arm, their motions are qualitatively similar to that of the robot.

B. Study 4: Human Deception

To see how humans would do deception in higher-dimensional spaces, we reproduced the study from Sec. II, but in the physical world: participants reached with their arms for objects on a table, and we recorded their trajectories using a motion capture system. We recruited 6 participants (3 male and 3 female, aged 18-41).

The strategies were indeed similar to the 2D case: 3 of the participants exaggerated the motion to the other object, then changed just before reaching it. Fig.9 shows one of the trajectories for the canonical scenario along with the end effector trace, which is qualitatively similar to the robot trajectory generated by the model: the hand first goes to the left, beyond the straight line connecting it to the target object, and then grazes past it to reach for the object on the right.

One of the participants adopted the animator strategy of aligning (in this case, the hand) with the other object first, and then moving straight toward it. Another participant used their torso more than their arm motion to indicate one goal or the other. A last participant used the strategy of moving predictably to the other goal (Fig.2e), bringing up a great point that in a game setting, exaggerating to convey intent would make the opponent suspicious that they are trying to deceive.

Overall, despite the diversity in approaches, the majority did match the model's output.

VI. Study 5: Implications of Deception for HRI

Our studies thus far test that the robot can generate deceptive motion. Our final study is about what effect this has on the perceptions and attitudes of people interacting with the robot.

Although no prior work has investigated deceptive *motion*, some studies have looked into deceptive robot *behavior* during games. A common pattern is that unless the behavior is very obviously deceptive, users tend to perceive being deceived as *unintentional*: an error on the side of the robot [23, 28, 32]. In a taxonomy of robot deception, Shim et al. [27] associate *physical* deception with unintentional, and *behavioral* deception with intentional. Deceptive motion could be thought of as either of the two, leading to our main question for this study:

Do people interpret deceptive motion as intentional?

And, if so, what implications does this have on how they perceive the robot? Literature on the ethics of deception cautions about a drop in trust [2, 20], while work investigating games with cheating robots measures an increase in engagement [28, 32]. We use these as part of our dependent measures in the study.

We also measure perceived intelligence, because deception is also associated with the agent having a theory of mind about the deceived [4].

A. Experimental Setup

Procedure. The participants play a game against the robot, in which they have to anticipate which bottle (of the two in front of them) the robot will grab, and steal it from the robot, like in Fig.10. The faster they do this, the higher their score in the game.

Before the actual game, in which the robot executes a deceptive trajectory, they play two practice rounds (one for each bottle) in which the robot moves predictably. These are meant to expose them to how the robot can move, and get them to form a first impression of the robot.

We chose to play two practice rounds instead of one for two reasons: (1) to avoid changing the participants' prior on what bottle is next, and (2) to show participants that the robot can move directly to either bottle, be it on the right or left. However, to still leave some suspicion about how the robot can move, we translate the bottles to a slightly different position for the deception round.

Dependent Measures. After the deception round, we first ask the participants whether the robot's motion made it seem (initially) like it was going to grab the other bottle. If they say yes, then we ask them whether they think that was intentional, and whether they think the robot is reasoning about what bottle they will think it would pick up (to test attribution of a theory of mind).

Both before and after the deception round, we ask participants to rate, on a 7 point Likert scale, how intelligent, trustworthy, engaging, and good at being an adversary the robot is.

Participants. We recruited 12 participants from the local community (9 male, 3 female, aged 20-44).

Hypothesis. *The ratings for intelligence, engagement, and adversary increase after deception, but trust drops.*

B. Analysis

The users' interpretation was surprisingly mixed, indicating that deception in motion can be subtle enough to be interpreted as accidental.

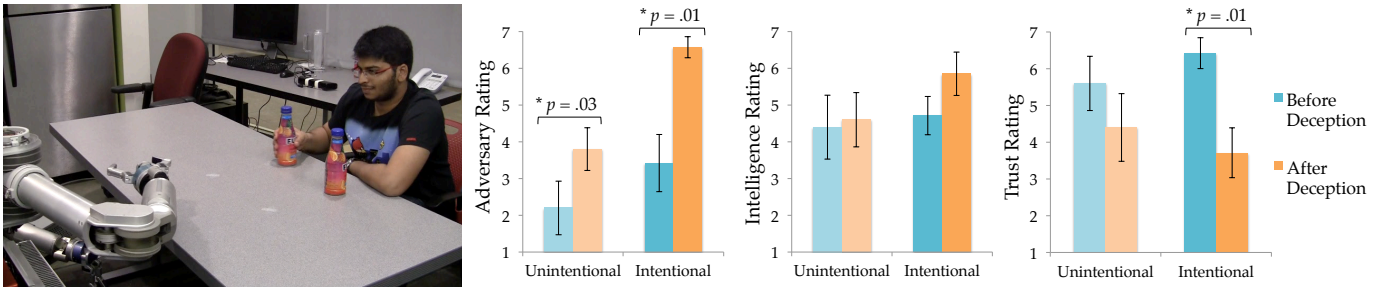

Fig. 10. A snapshot of the deception game, along with the adversary and trust ratings: after deception, users rate the robot's skill as an adversary higher, and trust in the robot decreases. The difference is larger when they perceive the deception as intentional.

Out of 12 users, 7 thought the robot was intentionally deceiving them, while 5 thought it was unintentional. Among those 5, 2 thought that the deceptive motion was hand-generated by a programmer, and not autonomously generated by the robot by reasoning about their inference. The other 3 attributed the way the motion looked to a necessity, rationalizing it based on how they thought the kinematics of the arm worked, e.g. "it went in that direction because it had to stretch its arm out".

Analyzing the data across all 12 users (Fig.10), the rating of the robot as an adversary increased significantly (paired t-test, $t(11) = 4.60$, $p < .001$), and so did the rating on how engaging the robot is ($t(11) = 2.45$, $p = .032$), while the robot's trustworthiness dropped ($t(11) = -3.42$, $p < .01$). The intelligence rating had a positive trend (increased by .75 on the scale), but it was not significant ($p = .11$). With Bonferroni corrections for multiple comparisons, only adversary and trust remain significant, possibly because of our small sample size. Further studies with larger sample sizes would be needed to investigate the full extent of the effect of deceptive motion on the interaction.

We also analyzed the data split by whether deception was perceived as intentional – this leads to even smaller sample sizes, meaning these findings are very preliminary and should be interpreted as such. We see larger differences in all metrics in the intentional case compared to the unintentional. This is somewhat expected: if deception is attributed to an accident, it is not a reflection on the robot's qualities. The exception is the rating of the robot as an adversary: both ratings increase significantly (Fig.10), perhaps because even when the deception was accidental, it was still effective at winning the game.

There was one user whose trust did not drop, despite finding deception intentional. He argued that the robot did nothing against the rules. Other users, however, commented that even though the robot played by the rules, they now know that it is capable of tricking them and thus trust it less.

VII. Discussion

In this work, we analyzed human strategies for deceptive motion, introduced a mathematical model that enables the robot to autonomously generate such motions, and tested users' reactions to being deceived.

Findings. We found that the model performs on par with the expert demonstration, and that a creative novice user's demonstration performs surprisingly well. We also showed that the model can generalize to manipulator arms, and that the output for a somewhat anthropomorphic arm is similar to human deceptive arm motion.

Finally, we found that users are mixed in perceiving the deceptive motion as intentional vs. unintentional, and that deception can increase ratings of engagement, intelligence, and adversarial standing, but can negatively impact trust. Even though the robot plays by the rules, the users become aware of its capability to deceive.

Future Directions. Although our analysis covered many aspects of deception throughout our five user studies, much remains unexplored. One of the richest areas of future work is long-term interactions. Our evaluation studies only test that users are deceived once, and our interaction study only tests the immediate user reaction to being deceived.

However, deceiving repeatedly would entail higher-level game-theoretic decisions, such as changing strategies. Similarly, people's reactions after being deceived repeatedly and in various situations would likely change as well: those who find it accidental orixginally would probably realize it to be intentional.

Deception has a counterpart in clear, intent-expressive communication, but comes with additional burdens, like the need to change strategies. For two goals, they have a symmetry: the easier it is to be legible, the more extra energy it takes to be deceptive. At the same time, deception has additional flexibility: the choice of which goal to convey. Depending on the scenario, some goals will allow for more convincing trajectories, and quickly finding the best such decoy remains a challenge.

Finally, our model showed that it can express different strategies, and our studies showed that the geometry of the path is important for deception. Although important, geometry is not everything. An area for further exploration is modeling more creative strategies, such as circling an object to express "hovering", or explicitly using timing (e.g. pausing to express doubt).

Overall, we are excited to have brought about a better understanding of deception through the *motion* channel, and look forward to exploring these remaining challenges in our future work.

References

[1] R. Alami, A. Clodic, V. Montreuil, E. A. Sisbot, and R. Chatila. Toward human-aware robot task planning. In *AAAI Spring Symposium*, pages 39–46, 2006.

[2] R. C. Arkin. The ethics of robotic deception. *The Computational Turn: Past, Present, Futures?*, 2011.

[3] M. Beetz, F. Stulp, P. Esden-Tempski, A. Fedrizzi, U. Klank, I. Kresse, A. Maldonado, and F. Ruiz. Generality

and legibility in mobile manipulation. *Autonomous Robots*, 28:21–44, 2010.

[4] C. Biever. Deceptive robots show theory of mind. *New Scientist*, 207(2779):24–25, 2010.

[5] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems (IROS)*, pages 708–713. IEEE, 2005.

[6] B. R. Brewer, R. L. Klatzky, and Y. Matsuoka. Visual-feedback distortion in a robotic rehabilitation environment. *Proceedings of the IEEE*, 94(9):1739–1751, 2006.

[7] G. Csibra and G. Gergely. Obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, 124(1):60 – 78, 2007.

[8] R. B. D'AGOSTINO. A second look at analysis of variance on dichotomous data. *Journal of Educational Measurement*, 8(4):327–333, 1971.

[9] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 2013.

[10] M. Dewar. *The art of deception in warfare*. David & Charles Publishers, 1989.

[11] A. Dragan, K. Lee, and S. Srinivasa. Legibility and predictability of robot motion. In *Human-Robot Interaction*, 2013.

[12] A. Dragan and S. Srinivasa. Formalizing assistive teleoperation. In *Robotics: Science and Systems*, July 2012.

[13] A. Dragan and S. Srinivasa. Generating legible motion. In *Robotics: Science and Systems*, June 2013.

[14] A. Dragan and S. Srinivasa. Familiarization to robot motion. In *Human-Robot Interaction*, 2014.

[15] D. Floreano, S. Mitri, S. Magnenat, and L. Keller. Evolutionary conditions for the emergence of communication in robots. *Current biology*, 17(6):514–519, 2007.

[16] R. Flynn. Anticipation and deception in squash. In *9th Squash Australia/PSCAA National Coaching conference*, 1996.

[17] G. Gergely, Z. Nadasdy, G. Csibra, and S. Biro. Taking the intentional stance at 12 months of age. *Cognition*, 56(2):165 – 193, 1995.

[18] M. Gielniak and A. Thomaz. Generating anticipation in robot motion. In *RO-MAN*, 2011.

[19] N. D. Goodman and A. Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184, 2013.

[20] P. Hancock, D. Billings, and K. Schaefer. Can you trust your robot? *Ergonomics in Design: The Quarterly of Human Factors Applications*, 19(3):24–29, 2011.

[21] R. C. Jackson, S. Warren, and B. Abernethy. Anticipation skill and susceptibility to deceptive movement. *Acta psychologica*, 123(3):355–371, 2006.

[22] T. S. Jim Mainprice, E. Akin Sisbot and R. Alami. Planning safe and legible hand-over motions for human-robot interaction. In *IARP Workshop on Technical Challenges for Dependable Robots in Human Environments*, 2010.

[23] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson. Do people hold a humanoid robot morally accountable for the harm it causes? In *International conference on Human-Robot Interaction*, pages 33–40, 2012.

[24] S. S. Raza Abidi, M. Williams, and B. Johnston. Human pointing as a robot directive. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 67–68. IEEE Press, 2013.

[25] E. Sato, T. Yamaguchi, and F. Harashima. Natural interface using pointing behavior for human–robot gestural interaction. *Industrial Electronics, IEEE Transactions on*, 54(2):1105–1112, 2007.

[26] J. Shim and R. C. Arkin. Biologically-inspired deceptive behavior for a robot. In *From Animals to Animats 12*, pages 401–411. Springer, 2012.

[27] J. Shim and R. C. Arkin. A taxonomy of robot deception and its benefits in hri. 2013.

[28] E. Short, J. Hart, M. Vu, and B. Scassellati. No fair!! an interaction with a cheating robot. In *International Conference on Human-Robot Interaction (HRI)*, pages 219–226, 2010.

[29] N. Smeeton and A. Williams. The role of movement exaggeration in the anticipation of deceptive soccer penalty kicks. *British Journal of Psychology*, 103(4):539–555, 2012.

[30] L. Takayama, D. Dooley, and W. Ju. Expressing thought: improving robot readability with animation principles. In *HRI*, 2011.

[31] K. Terada and A. Ito. Can a robot deceive humans? In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 191–192. IEEE, 2010.

[32] M. Vázquez, A. May, A. Steinfeld, and W.-H. Chen. A deceptive robot referee in a multiplayer gaming environment. In *Collaboration Technologies and Systems (CTS), 2011 International Conference on*, pages 204–211. IEEE, 2011.

[33] A. Vogel, C. Potts, and D. Jurafsky. Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[34] A. R. Wagner and R. C. Arkin. Robot deception: recognizing when a robot should deceive. In *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on*, pages 46–54. IEEE, 2009.

[35] A. R. Wagner and R. C. Arkin. Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics*, 3(1):5–26, 2011.

[36] B. Whaley. Toward a general theory of deception. *The Journal of Strategic Studies*, 5(1):178–192, 1982.

[37] T. Yamaguchi, E. Sato, and S. Sakurai. Recognizing pointing behavior using humatronics oriented human-robot interaction. In *33rd Annual Conference of the IEEE Industrial Electronics Society*, pages 4–9, 2007.

[38] M. Zucker, N. Ratliff, A. Dragan, M. Pivtoraiko, M. Klingensmith, C. Dellin, J. A. D. Bagnell, and S. Srinivasa. Chomp: Covariant hamiltonian optimization for motion planning. *International Journal of Robotics Research*, May 2013.