Auton Robot (2015) 39:331–345
DOI 10.1007/s10514-015-9458-8

Deceptive robot motion: synthesis, analysis and experiments

Anca Dragan[1] · **Rachel Holladay**[1] · **Siddhartha Srinivasa**[1]

Received: 24 November 2014 / Accepted: 2 July 2015 / Published online: 19 July 2015

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 studies on the implications of deceptive motion for human-robot interactions and the effects of iterated deception.

Keywords Deception · Motion planning · Human robot interaction · Legibility

1 Introduction

Much robotics research explores how robots can communicate effectively, via speech (Deits et al. 2013; Vogel et al. 2013; Goodman and Stuhlmüller 2013), gesture (Sato et al. 2007; Raza Abidi et al. 2013; Yamaguchi et al. 2007; Breazeal et al. 2005; Holladay et al. 2014), or motion (Takayama et al. 2011; Beetz et al. 2010; Alami et al. 2006; Jim Mainprice et al. 2010; Dragan and Srinivasa 2013; Gielniak and Thomaz 2011). 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 (Dewar 1989), but its uses go far beyond (Castelfranchi 2000; Shim and Arkin 2013, 2014; Nijholt 2010; Adar et al. 2013; Williams et al. 2014). At its core, deception conveys *intentionality* (Terada and Ito 2010), and that the robot has a *theory of mind* for the deceived (Biever 2010) which it can use to manipulate their beliefs. It makes interactions with robots more engaging, particularly during game scenarios (Vázquez et al. 2011; Terada and Ito 2010; Short et al. 2010).

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 (Flynn 1996), soccer (Smeeton and Williams 2012; Choudhury et al. 2011; Biswas et al. 2014) or rugby (Jackson et al. 2006). It can also find uses outside of competitions, such as tricking patients into exerting more force during physical therapy (Brewer et al. 2006). 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 seven user studies, from how humans would deceive, to how a robot can plan deceptive

This is one of several papers published in *Autonomous Robots* comprising the "Special Issue on Robotics Science and Systems".

✉ Rachel Holladay
rmh@andrew.cmu.edu

[1] The Robotics Institute, Carnegie Mellon University, Pittsburgh, USA

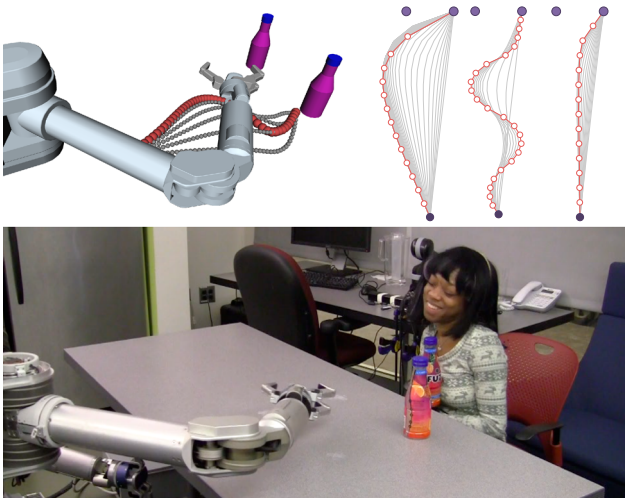


Fig. 1 *Top* deceptive motions produced by trajectory optimization. The trajectories on the *right* correspond to different strategies that humans adopt. *Bottom* a user's reaction when she first realizes the robot deceived her about which bottle it was going to grasp

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 (Sect. 2).

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.

We then cluster the demonstrations to reveal common strategies, and relate the emerging strategies to the theory of deceptive behavior in humans (Whaley 1982). 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 (Sect. 3), 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 (Wagner and Arkin 2009, 2011) or biology (Shim and Arkin 2012; Floreano et al. 2007; Arkin 2012).

Figure 1 (top) shows three examples generated by our model: (1) exaggerating: a trajectory that conveys the wrong goal (along with its higher-dimensional counterpart on the left), (2) switching: a trajectory that switches between conveying either goal, and (3) ambiguous: a trajectory 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 (Sect. 5). We find that all motions are significantly more deceptive than a baseline, and that the model performs almost equivalently to the animator trajectory.

We also compare the three strategies to verify that the strategy that conveys the wrong goal (exaggerating) is indeed more deceptive than the switching or ambiguous strategies from Fig. 1 (top), as predicted by the model.

4. Generalization: We show how our model generalizes to higher-DOF robots—manipulator arms (Sect. 6). 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. One of our studies explores the implications of the robot's deception through its motion to human-robot interaction (Sect. 7, 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.

6. Iterated deception: Our final study takes a first step towards analyzing deception in the long term, where a robot has to deceive not only once, but repeatedly. Here, we see that the most effective strategy for a one-time interaction is no longer effective when the robot has to deceive repeatedly, in an iterated game context.

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 (Arkin et al. 2012; Grodzinsky et al. 2015; Hancock et al. 2011; Reynolds and Ishikawa 2006; Sharkey 2015; Sharkey and Wood 2015). 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.

2 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 expres-

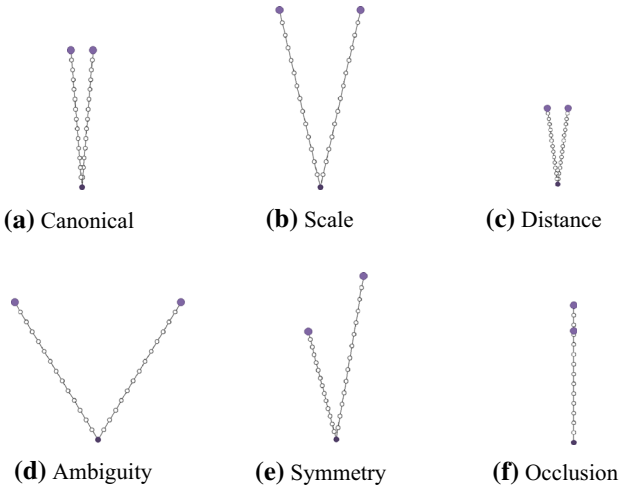


Fig. 2 Scenarios for study 1

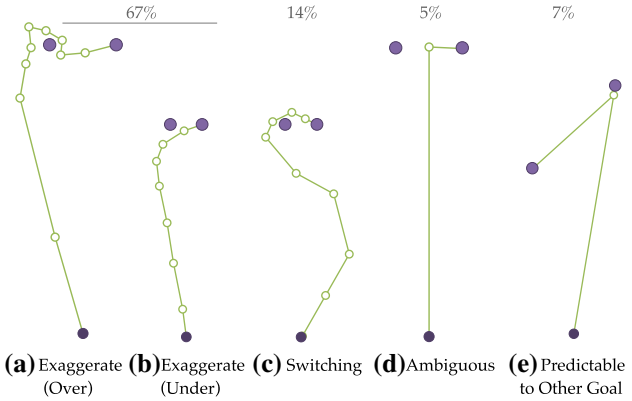


Fig. 3 User strategies for deception. The typical strategy exaggerates in the other direction and avoids the obstacle by going over it. A less common strategy of going under the obstacle closely matches the result of the model we use in Sect. 3, shown in Fig. 7 (*red*) (Color figure online)

sive motion for non-humans—to provide demonstrations of deceptive motion in various scenarios.

2.1 Experimental setup

Scenarios: We designed our scenarios by manipulating several variables. 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. This is the scenario from Fig. 2a.

To ensure that the strategies users produce are not biased by the particularities of this scenario, we test whether and how the trajectories change when different aspects of the environment change. From this canonical scenario (number 1), we manipulate:

(1) *goal side*, by changing which of the goals is the actual goal (number 2);
(2) *scale*, by scaling the entire scene (number 3);
(3) *distance*, by translating both goals vertically (number 4);
(4) *ambiguity*, by translating both goals horizontally (number 5);
(5) *goal symmetry*, by shifting one goal up and one down; we also tested both goals for this, as they were no longer symmetric and the results from number 2 would not generalize (numbers 6 and 7);
(6) *occlusion*, by lining one goal in front of the other; we again tested both goals for this because of asymmetry (numbers 8 and 9);

These were our main scenarios (Fig. 2). Additionally, we added *multiple goals* scenarios with three candidate goals instead of two, like in Fig. 4d, where we looked the the middle goal and one of the side goals (numbers 10 and 11). Thus, we had 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 Sect. 3).

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.

2.2 Analysis

Main scenarios: We started from the user comments, and identified 4 emerging strategies (one with 2 variations), shown in Fig. 3. We then classified each user trajectory as employing one of these 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 (Whaley 1982): it is a way of portraying false information (that the robot has a different goal from its actual goal) by offering a misleading alternate option.

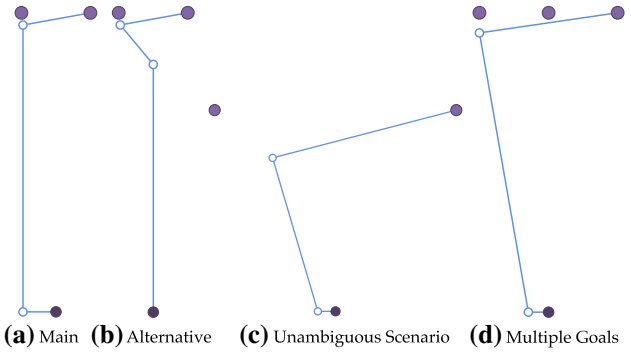


Fig. 4 Animator strategies

Among trajectories that follow this strategy, most (71 %) avoid the other goal by going over the top, like in Fig. 3a. 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. 3b.

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. 3c. This most closely resembles deception by *dazzling* (Whaley 1982), which is hiding the true by being confusing.

Approximately 5 % of trajectories were *ambiguous* (Fig. 3d), 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 (Whaley 1982), 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. 3e—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. 4 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. 4a).

The animator also proposed an alternative (Fig. 4b), 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. 4c, 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.

Trajectory timing: Some of the participants mentioned timing considerations when explaining their deceptive strategy.

Most participants thought that the robot should move quickly at the end of the exaggerating or ambiguous strategies, so that it can "dart to the goal before an observer could realize what was going on".

Other participants argued that fast movement conveys deliberate intent. Conversely, a slow moving robot conveys ambiguity, since the robot seems less sure of its path. The analogy referenced was that if one knows where they are headed, they walk quickly; however, when one is lost they tend to walk slowly and meander. Therefore, moving slowly can be used to convey uncertainty about the goal, which is a form of deception.

Ease of creation: We asked participants to rate, on a 7-point Likert scale, how easy it was to create the trajectories. A t test shows a significant effect for style of trajectory, $t(22) = 0.01$, $p < 0.0001$, with typical (predictable) trajectories rated as easier to create then deceptive trajectories. This is expected, since deception requires coming up with a strategy for communication, whereas typical movement only requires predicting what the character would normally do. In what follows, we introduce a mathematical model for deception that starts with a simpler model of predictable motion and builds on that to achieve communication.

3 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 (Dragan et al. 2013), 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 Dragan and Srinivasa (2012):

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

with Z a normalizer across the set of candidate goals $\mathcal{G}$ and $V_G(q) = \min_{\xi \in \Xi_{q \rightarrow G}} C[\xi]$. 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 (Dragan et al. 2013; Csibra and Gergely 2007).

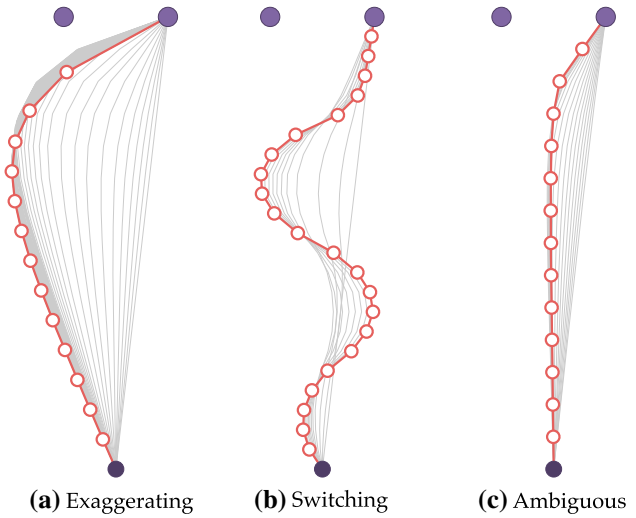
Given this, the robot can be most deceptive along the trajectory by minimizing the probability that the actual goal, G_{actual}, will be inferred:

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

Solution: We solve this trajectory optimization problem using functional gradient descent, following (Zucker et al. 2013; Dragan and Srinivasa 2013): we iteratively minimize the first order approximation of the objective plus a regularization term that keeps the trajectory from going to far from the approximation point. Here, distances between trajectories are measured with respect to a non-Euclidean inner product in the Hilbert space of trajectories, which propagates local gradient changes globally to the entire trajectory. This speeds up optimization and maintains smoothness. However, it also can introduce limitations: for instance, quick changes in velocity and direction could be useful for deception, as shown in the animator-produced demonstrations from Fig. 4. In future work, we plan to investigate the use of different inner products and their effects on deception.

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 (Zucker et al. 2013), and follow the derivation from Dragan and Srinivasa (2013) 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 (Zucker et al. 2013). This has been shown to match what users expect for a 2DOF robot (Dragan et al. 2013), and to have a high degree of predictability for a 7DOF robot arm as well (Dragan and Srinivasa 2014).



(a) Exaggerating **(b)** Switching **(c)** Ambiguous

Fig. 5 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

Our implementation parametrizes trajectories as vectors of a fixed number of configurations equally spaced in time.

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:

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

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

Solving this optimization problem leads to the trajectory in Fig. 5a, which qualitatively replicates the strategy in Fig. 3b of exaggerating the motion towards the other goal.

The *predictable-to-other-goal* strategy in Fig. 3e is similar, but instead of exaggerating, the robot moves predictably. However, prior work in conveying goals (Dragan et al. 2013) has shown exaggeration to be more effective.

The animator's main demonstration (Fig. 4a) 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 Sect. 5, we show that the model and animator trajectories perform similarly in practice.

The *switching* user trajectory (Fig. 3c) alternates between the goals. If $\sigma : [0, 1] \rightarrow \mathcal{G}$ 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:

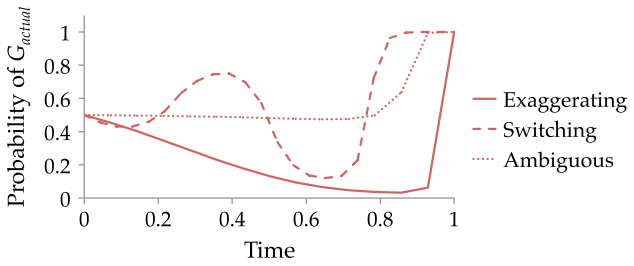


Fig. 6 The probability of the actual goal along each model trajectory

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

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. 5b, which alternates between conveying the goal on the right and the one on the left.

The *ambiguous* user trajectory (Fig. 3d) 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:

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

Figure 5c 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. 6 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.

4 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. 5, 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. 6), as well as according to the expert animator (see Sect. 2.2); we test this experimentally in our next study, which is specifically about comparing the strategies.

4.1 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. 3a, the main animator trajectory from Fig. 4a, the output of the model from Fig. 5a, 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. 7 (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 (i.e. a disc on the screen like the purple disc in Fig. 4) 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

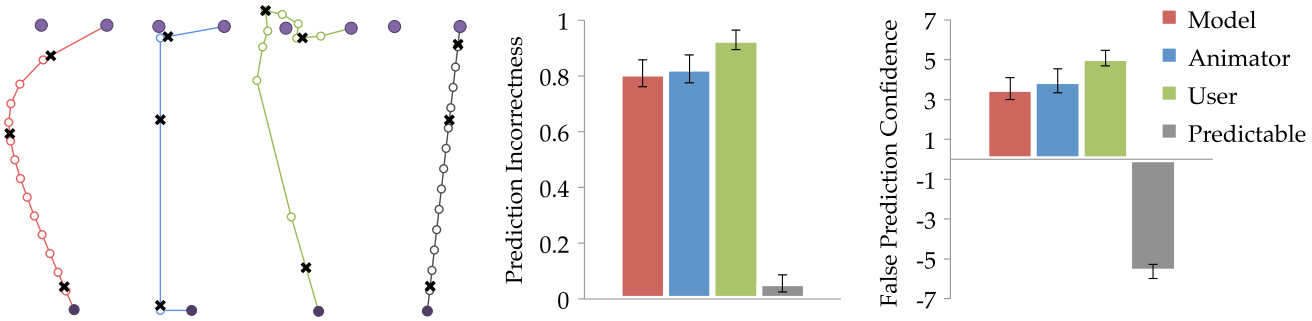


Fig. 7 The four trajectories: model, animator, user, and the predictable baseline, along with the comparison from our second user study (Color figure online)

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).

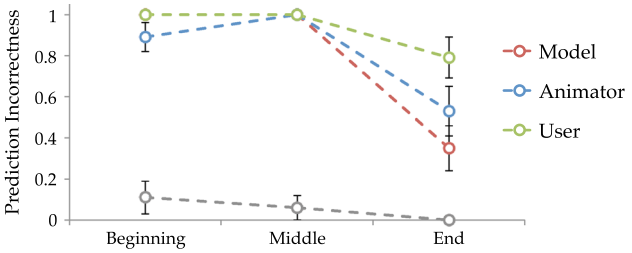


Fig. 8 A comparison of the four trajectories in terms of how deceptive they are across the three time points in study 2 (Color figure online)

4.2 Analysis

A factorial ANOVA on *incorrectness* [considered to be robust to dichotomous data (D'Agostino 1971)] 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 ($F(6, 222) = 5.5$, $p < .0001$) (Fig. 7).

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).

Figure 8 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 difference 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. 7 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.

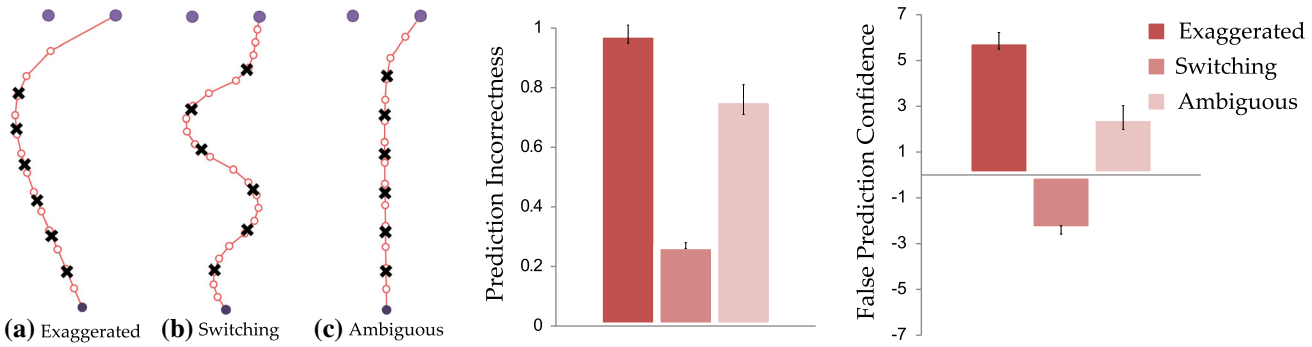


Fig. 9 A comparison among the three deception strategies in study 3: ambiguous, exaggerated and switching (Color figure online)

5 Study 3: comparing deception strategies

Our next study compares the effectiveness of the three deception strategies from Fig. 5: exaggerating, switching and ambiguous. From Fig. 6, we predict that exaggerating is more deceptive than the other two:

Hypothesis. *The exaggerating deceptive trajectory is more deceptive then the switching and ambiguous strategies.*

Manipulated factors and dependent measures: Similar to the previous study, we manipulated the *deception strategy* used (with the 3 levels outlined above), and the *time point* at which the trajectory is evaluated (with 6 time points equally spaced throughout the trajectory). This yielded a total of 18 conditions. We used the same dependent measures as in the second study, *incorrectness* and *false prediction confidence*.

Participants: We used a between-subjects design again, and recruited a total of 360 users (20 per condition) on Amazon's Mechanical Turk. We eliminated users who failed to answer a control question correctly, leading to 313 users (191 male, 122 female, aged 18–65).

5.1 Analysis

An ANOVA for *incorrectness* showed a significant main effect for *deception strategy* ($F(2, 310) = 77.98$, $p < .0001$), with the post-hoc revealing that all three strategies were significantly different from each other (all with $p < .0001$). An ANOVA for *false prediction confidence* yielded analogous findings.

As Fig. 9 shows, the exaggerating strategy was the most successful at deception, followed by the ambiguous strategy. This supports our hypothesis and the prediction of our model, since the exaggerating strategy assigns the lowest probability to the actual goal along the way (as shown in Fig. 6).

Figure 10 shows the correctness rate over time for the three strategies. This experimental evaluation has similar results to the theoretical prediction from Fig. 6: the exaggerating strategy decreases correctness over time, the switching strategy oscillates, and the ambiguous strategy stays closer to .5.

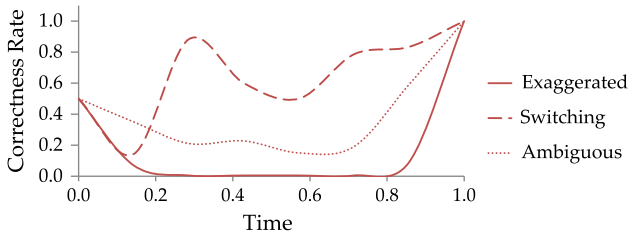


Fig. 10 The correctness rate for the three strategies as evaluated with users in study 3

However, we do observe differences from the predicted values. The exaggerating and ambiguous trajectories were more deceptive than expected, and the switching was less deceptive. In particular for switching, this could be an effect of the time point discretization we selected.

For the ambiguous strategy, many users were understandably unsure which object was the goal. However, a surprisingly large number of users were very confident that the robot was "clearly" moving towards one goal or the other, when in fact the motion was straight in the middle.

For the switching strategy, users commented on using one of two strategies to infer the goal. Some users based their choice on the robot's last movement, whether it tended to one side or the other. Other users noted, based on what section of the trajectory they saw, that the point robot spent more time on a certain side and therefore interpreted the point on that side to be the goal.

6 Generalization to arm motion

The previous section revealed that the mathematical model from Sect. 3 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. Note that this is in general done by defining the cost C over trajectories through the full configuration space, but C could also treat the end-effector or elbow trajectories separately. Investigating how different Cs could be used to obtain different deception strategies remains an area of future work.

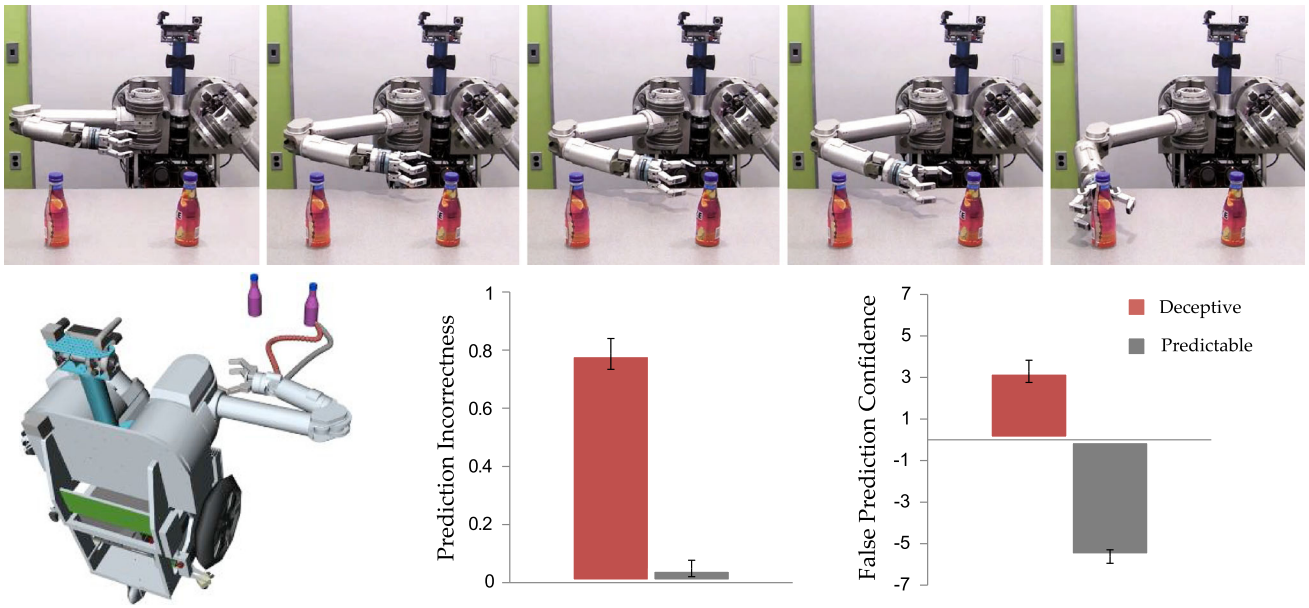


Fig. 11 *Top* the deceptive trajectory planned by the model. *Bottom* a comparison between this trajectory and the predictable baseline (Color figure online)

Figure 11 (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 Sect. 3. The difference is in the cost functional: the predictable trajectory minimizes C (Sect. 3), while the deceptive one minimizes the cost from (2). The planning time for either trajectory is based on CHOMP, and remains under a second.

Figure 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.

6.1 Study 4: robot trajectory evaluation

To evaluate whether this trajectory is really deceptive, we repeat our evaluation from Sect. 5, 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. 11. 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. 11 plots the results.

The users who were deceived relied on the principle of rational action (Gergely et al. 1995), 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".

Several users also perceived the motion to the bottle on the left as predictable, and thought it would be unpredictable if the robot were reaching for the bottle on the right, e.g. "If it were me, that's the one I would be going for". These statements imply that these users expected the robot's arm to function like that of a human.

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 a few users also quoted hand orientation, gaze and the elbow.

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 Sect. 3 can generalize to higher-dimensional spaces. Next, we run a further user study which indicates that this was no coincidence:

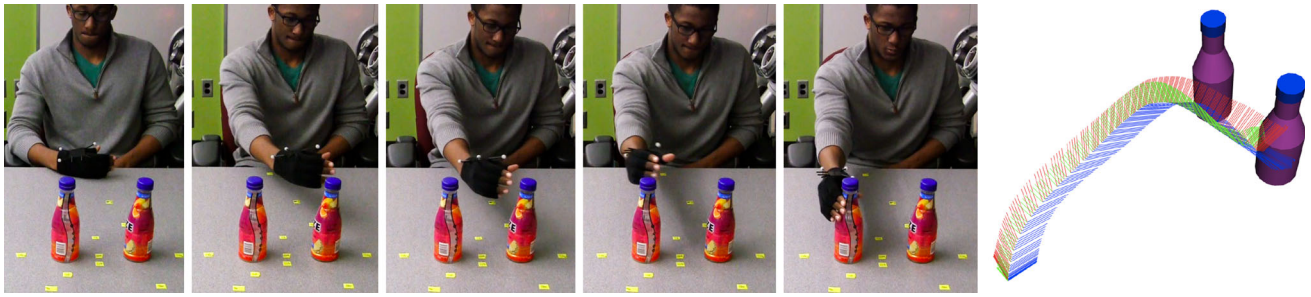


Fig. 12 A deceptive trajectory for a human arm from one of our participants, qualitatively similar to the robot trajectory from Fig. 11

when we ask humans to produce deceptive motion with their own arm, their motions are qualitatively similar to that of the robot.

6.2 Study 5: human deception

To see how humans would do deception in higher-dimensional spaces, we reproduced the study from Sect. 2, 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. 12 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. 3e), 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 seem to match the model's output.

7 Study 6: implications of deception for HRI

Our studies thus far test that the robot can generate deceptive motion. Our sixth 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 (Short et al. 2010; Vázquez et al. 2011; Kahn et al. 2012). In a taxonomy of robot deception, Shim and Arkin (2013) 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 (Hancock et al. 2011; Arkin 2011), while work investigating games with cheating robots measures an increase in engagement (Short et al. 2010; Vázquez et al. 2011). 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 (Biever 2010).

7.1 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. 13. 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.

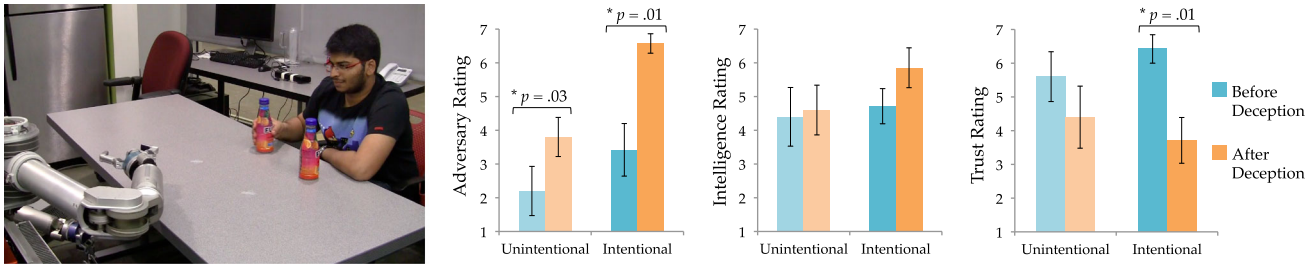


Fig. 13 A snapshot of the deception game, along with the adversary and trust ratings: after deception, users rate the robot's skill as an adversary higher, and trust in the robot decreases. The difference is larger when they perceive the deception as intentional (Color figure online)

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.*

7.2 Analysis

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

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. 13), 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. 13), 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.

8 Study 7: longer term effects

Thus far, our analysis of deception focuses on a single interaction with the user. In such a situation, we have seen that the exaggerated strategy is the most effective. Our final study is about deception in iterated (or repeated) interactions.

When the robot attempts to deceive repeatedly, we hypothesize that this decoy-like strategy will no longer be the most effective, because of user adaptation. Users will likely realize that the robot is always "lying" about its goal. Especially in two-goal situations, the actual goal will become clear: if the robot is always conveying the decoy goal, the actual goal must be the other one. Therefore, in such situations, an ambiguous strategy that holds off any information until the end might be more effective in the long term.

This study compares the exaggerated and ambiguous strategies in a context where the robot tries to deceive repeatedly. We leave out the switching strategy strictly for ease of evaluation, since the effectiveness of the switching strategy highly depends on the time point along the trajectory at which

we evaluate the inferred goal. Since this strategy is not overall more effective than the ambiguous, we focus on a different third strategy that explicitly accounts for the long term interaction.

We introduce a third, long-term strategy stemming from a game-theoretic argument. We formulate a game for iterated deception. At every interaction, the robot can choose to "lie" by conveying a decoy (the exaggerated deceptive strategy), or "be truthful" by conveying the actual goal (being legible Dragan and Srinivasa (2013), which in the two goal case optimizes the negative of the exaggerated utility function). Similarly, the human can choose to "trust" what the robot is conveying and infer that goal, or "distrust" and infer the other goal. In this game, the human gets utility 1 they infer the actual goal (and the robot −1), and the −1 otherwise (with the robot getting 1).

This is then a zero-sum game, with utility matrix:

	(R) be truthful	(R) lie
(R) trust	(H 1, R −1)	(H −1, R 1)
(R) distrust	(H −1, R 1)	(H 1, R −1)

The optimal column player (robot, R) strategy is the mixed strategy which assigns "lie" probability .5, and "be truthful" probability .5. Therefore, our third strategy uses the exaggerated deception in the first interaction, when the user does not expect deception, and employs this mixed strategy afterwards (Fig. 14).

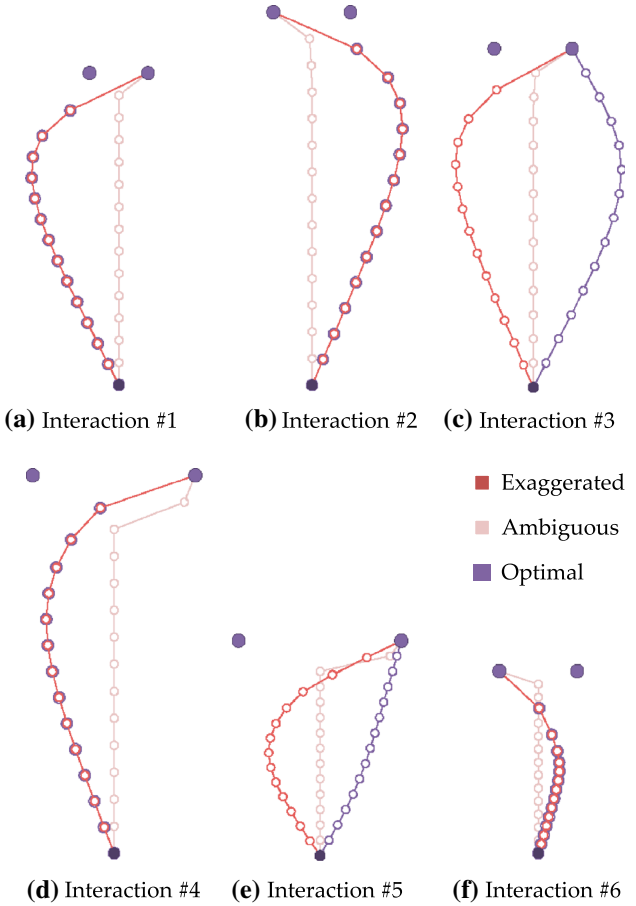


Fig. 14 The 6 scenarios for Study 7. The *dark and light red* and the exaggerated and ambiguous strategies, respectively. The *purple* trajectories stem from the optimal strategy, which is the same as the exaggerated one for the first interaction, and then follows the uniform mixed strategy which mixes deceiving and being legible (Color figure online)

8.1 Experimental design

Manipulated variables: We manipulate the *deception strategy* with 3 levels: exaggerated, ambiguous, and optimal.
Procedure: We generated trajectories for the point robot for 6 different scenarios, following each strategy. Fig. 14 shows the scenarios and the resulting trajectories. Each user therefore has 6 interactions with the robot where the robot attempts to deceive them. We choose the midpoint of the trajectory for evaluation, ask users for their goal inference, but then show them the remainder of the trajectory so that they can see what the actual goal was.
Dependent measures: We use the same measures as before: *incorrectness* and *false prediction confidence*.
Hypotheses.

H1 *The exaggerated and optimal strategies are more deceptive than the ambiguous for the first interaction.*

This is based on the results from Study 3.

H2 *The ambiguous and optimal strategies are more deceptive than the exaggerated strategy for the last interaction.*

If users adapt, they will no longer be deceived by the exaggerated strategy, whereas the ambiguous and optimal strategies will still deceive approximately 50 % of the users.

H3 *Overall, the optimal strategy is better than both the exaggerated and ambiguous strategies.*

We expect this because the optimal strategy should deceive in the beginning, and at some point converge to have a 50 % chance at deception. We would anticipate, however, that for very long interactions, the difference between optimal and ambiguous would become negligible.
Subject allocation: We used a between-subjects design, and recruited a total of 60 users (20 per condition) on Amazon's Mechanical Turk. We eliminated users who failed to answer a control question correctly, leading to 51 users (27 male, 24 female, aged 18–65).

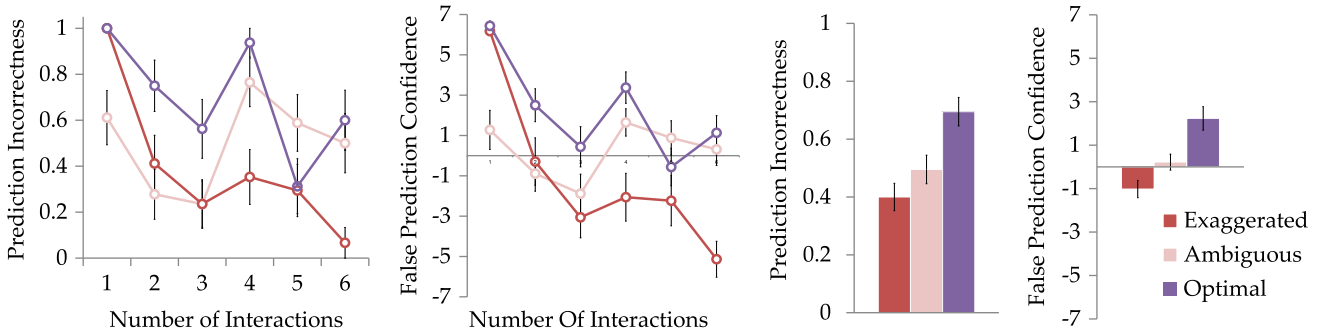


Fig. 15 Our deception measures for each interaction in the *two plots on the left*, and aggregated over all interactions in the two on the *right*. The exaggerated strategy does better than the ambiguous in the beginning, but degrades in the long term, making the ambiguous and optimal strategies more preferable for iterated deception (Color figure online)

8.2 Analysis

A repeated-measures ANOVA for *incorrectness* using the number of interactions as a covariate did show a significant effect for *deception strategy* ($F(2, 295) = 11.92$, $p < .0001$). Supporting **H3**, the post-hoc Tukey HSD revealed that the optimal strategy led to significantly more incorrect answers than both other strategies ($p < .0001$ for exaggerated and $p < .001$ for ambiguous). A repeated-measures ANOVA for *false prediction confidence* yielded analogous results.

Figure 15 plots the results. As expected (third plot), the ambiguous strategy aggregated an overall success rate of approximately 50 %, with the exaggerated strategy being worse and then optimal strategy being better.

As predicted by **H1**, the exaggerated and optimal strategies are more deceptive than the ambiguous strategy in the very beginning, for the first interaction. Users were unanimously deceived by the first two, and only deceived about 40 % of the time with the ambiguous strategy.

In line with **H2**, the ambiguous and optimal strategy tend to be more deceptive once the user and the robot have interacted repeatedly. For the final interaction, the ambiguous strategy was indeed significantly better than the exaggerated one, as indicated by a planned contrast based on **H2** ($F(1, 39) = 7.74$, $p < .01$). Across all interactions, however, there is not a significant difference between ambiguous and exaggerated. We believe this might be due to the limited number of users and the relatively small number of interactions, and not due to an inherent lack of ability of users to adapt to the exaggerated strategy and start making correct predictions. This is supported by decreasing trend of the exaggerated strategy across iterations.

With the optimal strategy, which is sometimes deceptive and sometimes legible, we see that users tend to be more deceived in interactions that use the deception (number 1, 2, 4, and 6), and less in interactions that use legibility (numbers 3 and 5). Unlike the exaggerated condition however, for which the effectiveness severely drops by the 6th interaction (suggesting that users start "distrusting" the robot), users in the optimal condition still predict the goal conveyed by the robot, i.e. "trust" the robot. We expect that as the number of interaction increases, the percentage of users that would choose to "trust" the robot would become 50 %.

Overall, we see that in the long term the ambiguous strategy might be preferable to the exaggerated one, as it does not convey either goal. The same goes for randomly selecting between deceiving using the exaggerated strategy and being legible. For such a strategy, users keep trusting the robot's indication (as they see it sometimes be legible) for the fist few interactions, making it more effective even than the ambiguous one. We would however expect that over a much larger number of interactions, the two would be indistinguishable.

9 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.

With respect to reactions, we found that users are mixed in perceiving the deceptive motion as intentional vs. unintentional. Almost half the users thought that the robot was not purposefully deceiving them, and a quarter of the users essentially came up with excuses for why the robot had to move the way it did.

Across all user reactions, we found that deception significantly increases 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. These effects seem to be larger when users perceive the deception as intentional, but this last statement is a hypothesis that requires more testing in future work, with larger sample sizes.

We also saw that in the long term, the exaggerated strategy is no longer the most effective, as users start better anticipating that the robot will deceive. Instead, the ambiguous strategy, or a strategy that mixes conveying a decoy goal with conveying the true goal, are more useful.

Future directions: 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.

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).

Finally, goals are not the only type of intent that robots need to convey or deceive about. Properties of the motion, or higher level behaviors, would also be useful.

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

Adar, E., Tan, D.S., & Teevan, J. (2013). Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*, (pp. 1863–1872), ACM.

Alami, R., Clodic, A., Montreuil, V., Sisbot, E.A., & Chatila, R. (2006). Toward human-aware robot task planning. *AAAI Spring Symposium* (pp. 39–46).

Arkin, R.C. (2011). The ethics of robotic deception. *The computational turn: past, present, futures?*.

Arkin, R. C. (2012). Robots that need to mislead: Biologically-inspired machine deception. *IEEE Intelligent Systems*.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571–589.

Beetz, M., Stulp, F., Esden-Tempski, P., Fedrizzi, A., Klank, U., Kresse, I., et al. (2010). Generality and legibility in mobile manipulation. *Autonomous Robots*, *28*, 21–44.

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

Biswas, J., Mendoza, J. P., Zhu, D., Choi, B., Klee, S., & Veloso, M. (2014). Opponent-driven planning and execution for pass, attack, and defense in a multi-robot soccer team. *International conference on autonomous agents and multi-agent systems (AAMAS)* 2014.

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. *Intelligent robots and systems (IROS)*, (pp. 708–713), IEEE .

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

Castelfranchi, C. (2000). Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, *2*(2), 113–119.

Choudhury, S., Deb, A. K., & Mukherjee, J. (2011) Designing deception in adversarial reinforcement learning.

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

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

Deits, R., Tellex, S., Thaker, P., Simeonov, D., Kollar, T., & Roy, N. (2013). Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, *2*, 58–79.

Dewar, M. (1989). *The art of deception in warfare*. Newton Abbot: David & Charles Publishers.

Dragan, A., Lee, K., & Srinivasa, S. (2013). Legibility and predictability of robot motion. In *Human-Robot Interaction*.

Dragan, A., & Srinivasa, S. (2012). Formalizing assistive teleoperation. In *Robotics: Science and Systems*. Cambridge, MA: MIT Press.

Dragan, A., & Srinivasa, S. (2013). *Generating legible motion*. In Robotics: Science and Systems.

Dragan, A.,& Srinivasa, S. (2014) Familiarization to robot motion. In *Human-Robot Interaction*.

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

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

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

Gielniak, M., & Thomaz, A. (2011). Generating anticipation in robot motion. In *RO-MAN*.

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

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2015). Developing automated deceptions and the impact on trust. *Philosophy & Technology*, *28*(1), 91–105.

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

Holladay, R., Dragan, A., & Srinivasa, S.S. (2014). Legible robot pointing.

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

Jim Mainprice, T.S., Akin Sisbot, E., & Alami, R. (2010). Planning safe and legible hand-over motions for human-robot interaction. In *IARP Workshop on technical challenges for dependable robots in human environments*.

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

Nijholt, A. (2010) Computational deception.

Raza Abidi, S. S., Williams, M., & Johnston, B. (2013). Human pointing as a robot directive. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pp. 67–68. IEEE Press.

Reynolds, C., & Ishikawa, M. (2006). Robot trickery. In: *International workshop on ethics of human interaction with robotic, bionic, and AI systems: Concepts and policies*.

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

Sharkey, A. Dignity, older people, and robots.

Sharkey, A., & Wood, N. (2015). The paro seal robot: demeaning or enabling?.

Shim, J., & Arkin, R. C. (2012) Biologically-inspired deceptive behavior for a robot. In *From Animals to Animats 12*, pp. 401–411. Springer.

Shim, J.,& Arkin, R. C. (2013). A taxonomy of robot deception and its benefits in hri.

Shim, J.,& Arkin, R. C. (2014) Other-oriented robot deception: A computational approach for deceptive action generation to benefit the mark.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! an interaction with a cheating robot. In *International conference on Human-robot interaction (HRI)* (pp. 219–226).

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

Takayama, L., Dooley, D., & Ju, W. (2011). Expressing thought: improving robot readability with animation principles. In *HRI*.

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

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

Vogel, A., Potts, C., & Jurafsky, D. 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.

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

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

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

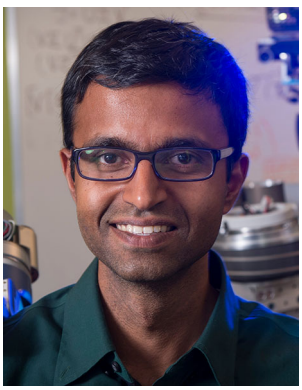
Williams, M.-A., Wang, X., Parajuli, P., Abedi, S., Youssef, M., & Wang, W. (2014). The fugitive: a robot in the wild. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, (pp. 111–111) ACM.

Yamaguchi, T., Sato, E., & Sakurai, S. (2007). Recognizing pointing behavior using humatronics oriented human-robot interaction. In *33rd annual conference of the IEEE industrial electronics society*, (pp. 4–9).

Zucker, M., Ratliff, N., Dragan, A., Pivtoraiko, M., Klingensmith, M., Dellin, C., et al. (2013). Covariant hamiltonian optimization for motion planning. *International Journal of Robotics Research*, *32*, 1164–1193.

Anca Dragan is a Ph.D. candidate at Carnegie Mellon's Robotics Institute, and a member of the Personal Robotics Lab. She was born in Romania and received her B.Sc. in Computer Science from Jacobs University Bremen in 2009. Her research lies at the intersection of robotics, machine learning, and human-robot interaction: she is interested in enabling robots to seamlessly work with, around, and in support of people. Anca is an Intel Ph.D. Fellow, a Siebel Scholar for 2015, a Dan David Prize Scholar for 2014, and a Google Anita Borg Scholar for 2012, and serves as General Chair in the Quality of Life Technology Center's student council.

Rachel Holladay is a second year undergraduate at Carnegie Mellon and a member of the Personal Robotics Lab. She is double majoring in Computer Science and Robotics. Her research focuses on manipulation and human-robot interaction, specifically enabling robots to be intent expressive and creating motion primitives for robot gestures.



Siddhartha Srinivasa is the Finmeccanica Associate Professor at Carnegie Mellon's Robotics Institute. His research goal is to to enable robots to robustly and gracefully interact with the world to perform complex manipulation tasks under uncertainty and clutter, with and around people. He founded and directs the Personal Robotics Lab. His research has received numerous awards, including 10 best paper award nominations. He is an Editor of the International Journal of Robotics Research, and IEEE IROS.