Legibility and Predictability of Robot Motion

Anca D. Dragan
Carnegie Mellon University

Kenton C.T. Lee
University of Pennsylvania

Siddhartha S. Srinivasa
Carnegie Mellon University

***Abstract*—A key requirement for seamless human-robot collaboration is for the robot to make its intentions clear to its human collaborator. A collaborative robot's motion must be *legible*, or intent-expressive. Legibility is often described in the literature as and effect of predictable, unsurprising, or expected motion. Our central insight is that predictability and legibility are fundamentally different and often contradictory properties of motion. We develop a formalism to mathematically define and distinguish predictability and legibility of motion. We formalize the two based on inferences between trajectories and goals in *opposing* directions, drawing the analogy to action interpretation in psychology. We then propose mathematical models for these inferences based on optimizing cost, drawing the analogy to the principle of rational action. Our experiments validate our formalism's prediction that predictability and legibility can contradict, and provide support for our models. Our findings indicate that for robots to seamlessly collaborate with humans, they must change the way they plan their motion.**

***Keywords*—human-robot collaboration, motion planning, trajectory optimization, formalism, manipulation, action interpretation**

I. Introduction

In this paper, we explore the problem where a robot and a human are working side by side to perform a tightly coupled physical task together, like clearing a table (Fig.1, and a running example in our paper).

The task amplifies the burden on the robot's motion: it must move in such a way that the human trusts and understands it. In robotics and animation, this is often achieved by *predictable* motion, that is *expected* – not surprising to a human, safe [1] or stereotypical [2].

However, the robot is also faced with another, often more critical burden of conveying its intent [3], e.g. which of the two bottles it is going to pick up to clean in Fig.1. In robotics and animation, this is often achieved by *legible* motion, that is *intent-expressive* – it enables the inference of intentions [4], it is "readable" [5], "anticipatory" [6], or "understandable" [7].

Predictable and legible motion can be correlated. For example, in an unambiguous situation, where an actor's observed motion matches what is expected for a given intent (i.e. is predictable), then this intent can be used to explain the motion. If this is the only intent which explains the motion, the observer can immediately infer the actor's intent, meaning that the motion is also legible. As a consequence, predictability and legibility are often treated as an inseparable couple of desirable properties of robot motion [1], [2], [8]–[10].

The writing domain, however, clear distinguishes the two. The word *legibility*, traditionally an attribute of written text [11], refers to the quality of being easy to read. When we write legibly, we try consciously, and with some effort, to make our writing clear and readable to someone else, like in Fig.1(top, right). The word *predictability*, on the other hand, refers to the quality of matching expectation. When we write predictably, we fall back to old habits, and write with minimal effort, as in Fig.1(top, left).

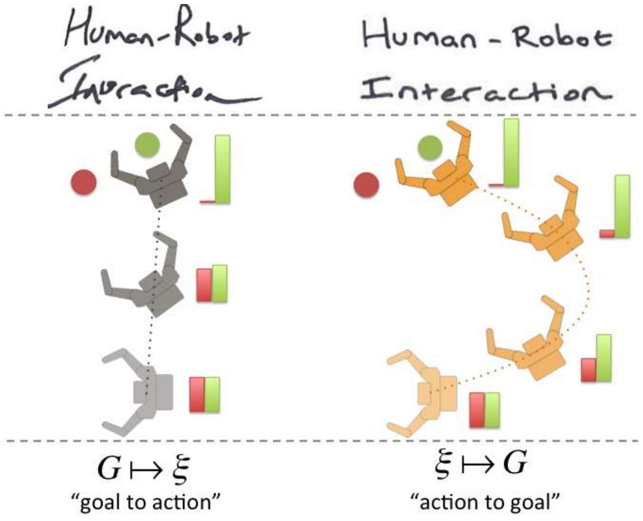


Fig. 1. Above: Predictable, day-to-day, expected handwriting vs. legible handwriting. Center: A predictable and a legible trajectory of a robot's hand for the same task of grasping the green object. Below: Predictability and legibility stem from inferences in opposing directions.

As a consequence, our legible and predictable writings are *different*: our friends do not expect to open our diary and see our legible writing style. They rightfully assume the diary will be written for us, and expect our usual, day-to-day style.

In this paper, we show that legibility and predictability are *different* in motion as well. Our main contribution is a formalism that emphasizes this difference, showing that the two properties stem from inferences in *opposing* directions (Fig.1,below): expressing intent means enabling an observer to infer the goal of the motion (an inference from a trajectory to a goal), while matching expectation means matching the motion inferred by an observer based on knowledge of the goal (an inference from a goal to a trajectory). This opposition leads to our central insight:

> *Predictability and legibility are fundamentally different and often contradictory properties of motion.*

Ambiguous situations, occurring often in daily tasks, make this opposition clear: more than one possible intent can be used to explain the motion observed so far, rendering the predictable motion illegible. Fig.1(center) exemplifies the effect of this contradiction. The robot hand's motion on the left is predictable in that it matches expected behavior. The hand reaches out directly towards the target. But, it is not legible, failing to make the intent of grasping the green object clear. In contrast, the trajectory on the right is more legible, making it clear that the target is the green object by deliberately bending away from the red object. But it is less predictable, as it does not match the expected behavior of reaching directly. We will show in Sections III and IV how we can quantify this effect with Bayesian inference, which allows us to derive, among other things, the online probabilites of the motion reaching for either object, illustrated as bar graphs in Fig.1.

Our work makes the following three contributions:

1. We formalize legibility and predictability in the context of goal-directed motion in Section II as stemming from inferences in *opposing* directions. The formalism emphasizes their difference, and directly relates to the theory of action interpretation [12] and the concepts of "action-to-goal" and "goal-to-action" inference. Our formalism also unifies previous descriptions of legibility, quantifying readability and understandability, and encouraging anticipation as a direct consequence of our definitions.

2. Armed with mathematical definitions of legibility and predictability, we propose a way in which a robot could model these inferences in order to evaluate and generate motion that is legible or predictable (Sections III and IV). The models are based on cost optimization, resonate with the principle of rational action [13], [14], and echo earlier works on action understanding via inverse planning [15].

3. We demonstrate that legibility and predictability are contradictory not just in theory, but also in practice. We present an extensive experiment for three characters that differ in their complexity and anthropomorphism: a simulated point robot, the bi-manual mobile manipulator HERB [16], and a human (Section V). The experiment confirms the contradiction between predictable and legible motion, and reveals interesting challenges (Section VI). We found, for instance, that different people expect a complex robot like HERB to act in different ways: for a robot to be predictable, it must adapt to the particulars of the observer.

The difference between legibility and predictability of motion is crucial for human-robot interaction, in particular for collaboration between humans and robots. Collaboration is a delicate dance of prediction and action, where agents must predict their collaborator's intentions as well as make their own intentions clear – they must act legibly. We are excited to be taking an essential step towards better human-robot collaboration: by emphasizing the difference between legibility and predictability, we advocate for a different approach to motion planning, in which robots decide between optimizing for legibility and optimizing for predictability, depending on the context they are in.

II. Formalizing Legibility and Predictability

So far, we have identified that legible motion is intent-expressive, and predictable motion matches what is expected. Here, we formalize these definitions for the context of goal-directed motion, where a human or robot is executing a trajectory towards one goal G from a set of possible goals $\mathcal{G}$, like in Fig.1. In this context, G is central to both properties:

Definition 2.1: Legible motion is motion that enables an observer to quickly and confidently infer the correct goal G.

Definition 2.2: Predictable motion is motion that matches what an observer would expect, given the goal G.

A. Formalism

1) Legibility: Imagine an observer watching the orange trajectory from Fig.1. As the robot's hand departs the starting configuration and moves along the trajectory, the observer is running an inference, predicting which of the two goals it is reaching for. We denote this inference function that maps (snippets of) trajectories from all trajectories Ξ to goals as

$$\mathcal{I}_L : \Xi \rightarrow \mathcal{G}$$

The bar graphs next to the hands in Fig.1 signify the observer's predictions of the two likely goals. At the very beginning, the trajectory is confusing and the observer has little *confidence* in the inference. However, the observer becomes confident very *quickly* – even from the second configuration of the hand along the trajectory, it becomes clear that the green object is the target. This quick and confident inference is the hallmark of legibility.

We thus formalize *legible* motion as motion that enables an observer to *confidently* infer the *correct* goal configuration G after observing only a snippet of the trajectory, $\xi_{S\rightarrow Q}$, from the start S to the configuration at a time t, $Q = \xi(t)$:

$$\mathcal{I}_L(\xi_{S\rightarrow Q}) = G$$

The *quicker* this happens (i.e. the smaller t is), the more legible the trajectory is.

This formalizes terms like "readable" [5], or "understandable" [7], and encourages "anticipatory" motion [6] because it brings the relevant information for goal prediction towards the beginning of the trajectory, thus lowering t. The formalism can also generalize to outcome-directed motion (e.g. gestures such as pointing at, waving at, etc.) by replacing the notion of goal with that of an outcome – here, legible motion becomes motion that enables quick and confident inference of the desired outcome.

2) Predictability: Now imagine someone knowing that the hand is reaching towards the green goal. Even before the robot has moved, the observer creates an expectation, making an inference on how the hand will move – for example, that the hand will start turning towards the green object as it is moving directly towards it. We denote this inference function mapping goals to trajectories as

$$\mathcal{I}_P : \mathcal{G} \rightarrow \Xi$$

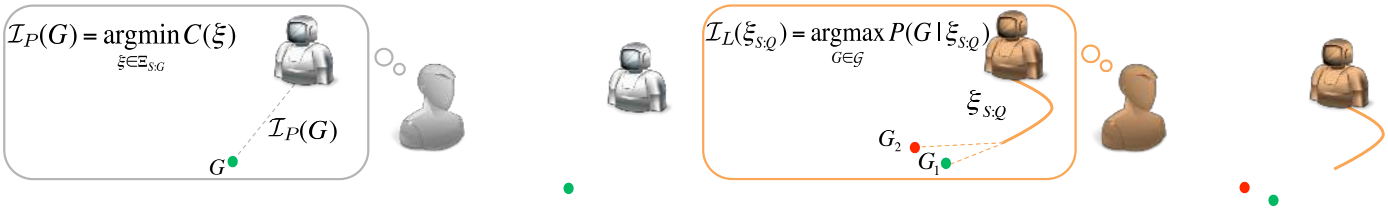
We formalize *predictable* motion as motion for which the trajectory $\xi_{S\rightarrow G}$ matches this inference:

$$\mathcal{I}_P(G) = \xi_{S\rightarrow G}$$

The better the actual trajectory matches the inference, measurable for example using a distance metric between $\mathcal{I}_P(G)$ and $\xi_{S\rightarrow G}$, the more predictable the trajectory is.

B. Connection to Psychology

A growing amount of research in psychology suggests that humans interpret observed behaviors as goal-directed actions [12], [17]–[21], a result stemming from studies observing infants and how they show surprise when exposed to inexplicable



$\mathcal{I}_P(G) = \underset{\xi\in\Xi_{S:G}}{\text{argmin}}\, C(\xi)$
$\mathcal{I}_P(G)$
G
$\mathcal{I}_L(\xi_{S:Q}) = \underset{G\in\mathcal{G}}{\text{argmax}}\, P(G\,|\,\xi_{S:Q})$
$\xi_{S:Q}$
G_2
G_1

Fig. 2. Our models for $\mathcal{I}_P$ and $\mathcal{I}_L$: the observer expects the robot's motion to optimize a cost function C ($\mathcal{I}_P$, left), and identifies based on C which goal is most probable given the robot's motion so far ($\mathcal{I}_L$, right).

action-goal pairings. Csibra and Gergeley [12] summarize two types of inference stemming from the interpretation of actions as goal directed: "action-to-goal" and "goal-to-action".

"Action-to-goal" refers to an observer's ability to infer someone's goal state from their ongoing actions (e.g. because they are pouring coffee beans into the grinder, the will eventually hold a cup of coffee). "Action-to-goal" inference answers the question "What is the function of this action?".

"Goal-to-action" refers to an observer's ability to predict the actions that someone will take based on their goal (e.g. because they want to make coffee, they will will pour coffee beans into the grinder). "Goal-to-action" inference answers the question "What action would achieve this goal?".

This has a natural connection to our formalism. In goal-directed motion, actions are trajectories and goals are goal configurations. Thus the inference occurring in legibility, from trajectory to goal, $\xi_{S\to Q} \mapsto G$, relates naturally to "action-to-goal" inference. Likewise, the inference occurring in predictability, from goal to trajectory, $G \mapsto \xi_{S\to G}$, relates naturally to "goal-to-action".

C. Summary

Our formalism emphasizes the difference between legibility and predictability in theory: they stem from inferences in *opposing directions* (from trajectories to goals vs. from goals to trajectories), with strong parallels in the theory of action interpretation. In what follows, we introduce one way for a robot to model these two inferences (summarized in Fig.2), and present an experiment that emphasizes the difference between the two properties in practice.

III. Modeling Predictable Motion

A. The Trajectory Inference $\mathcal{I}_P$

To model $\mathcal{I}_P$ is to model the observer's expectation. One way the robot could do so is by assuming that the human observer expects it to be a rational agent acting efficiently [12] or justifiably [14] to achieve a goal. This is known as the principle of rational action [13], [14], and it has been shown to apply to non-human agents, including robots [22]. The robot could model this notion of "efficiency" via a cost function defining what it means to be efficient. For example, if the observer expected the robot's hand to move directly towards the object it wants to grasp (as opposed to taking an unnecessarily long path to it), then "efficiency" would be defined by the cost function penalizing the trajectory's length.

Throughout this paper, we will refer to the cost function modeling the observer's expectation as C:

$$C : \Xi \to \mathbb{R}^+$$

with lower costs signifying more "efficient" trajectories.

The most predictable trajectory is then the most "efficient":

$$\mathcal{I}_P(G) = \arg \min_{\xi\in\Xi_{S\to G}} C(\xi) \qquad (1)$$

C represents what the observer expects the robot to optimize, and therefore encompasses every aspect of the observer's expectation, including (when available) body motion, hand motion, arm motion, and gaze.

B. Evaluating and Generating Predictability

Predictability can be evaluated based on C: the lower the cost, the more predictable (expected) the trajectory. We propose a predictability *score* normalized from 0 to 1:

$$\mathsf{predictability}(\xi) = \exp\big(-C(\xi)\big) \qquad (2)$$

Generating predictable motion means maximizing this score, or equivalently minimizing the cost function C – as in (1). This presents two major challenges: learning C, and minimizing C.

First, the robot needs access to the cost function C that captures how the human observer expects it to move. If the human observer expects human-like motion, animation (e.g. [23]–[25]) or biomechanics (e.g. [26], [27]) literature can serve to provide approximations for C. Our experiment (Section V) uses trajectory length as a proxy for the real C, resulting in the shortest path to goal – but this is merely one aspect of expected behavior. As our experiment will reveal, efficiency of robot motion has different meanings for different observers. If the observer were willing to provide examples of what they expect, the robot could learn how to act via Learning from Demonstration [28]–[30] or Inverse Reinforcement Learning [31]–[33]. Doing so in a high-dimensional space, however, is still an active area of research.

Second, the robot must find a trajectory that minimizes C. This is tractable in low-dimensional spaces, or if C is convex. While efficient trajectory optimization techniques do exist for high-dimensional spaces and non-convex costs [34], they are subject to local minima, and how to alleviate this issue in practice remains an open research question [35], [36].

IV. Modeling Legible Motion

A. The Goal Inference $\mathcal{I}_L$

To model $\mathcal{I}_L$ is to model how the observer infers the goal from a snippet of the trajectory $\xi_{S\to Q}$. One way to do so is by assuming that the observer compares the possible goals in the scene in terms of how probable each is given $\xi_{S\to Q}$. This is supported by action interpretation: Csibra and Gergeley [12] argue, based on the principle of rational action, that humans assess which end state would be most efficiently brought about by the observed ongoing action.Taking trajectory length again as an example for the observer's expectation, this translates to predicting a goal because $\xi_{S\to Q}$ moves directly toward it and away from the other goals, making them less probable.

One model for $\mathcal{I}_L$ is to compute the probability for each goal candidate G and to choose the most likely:

$$\mathcal{I}_L(\xi_{S\to Q}) = \arg\max_{G\in\mathcal{G}} P(G|\xi_{S\to Q}) \tag{3}$$

To compute this probability, we start with Bayes' Rule:

$$P(G|\xi_{S\to Q}) \propto P(\xi_{S\to Q}|G)P(G) \tag{4}$$

where $P(G)$ is a prior on the goals which can be uniform in the absence of prior knowledge, and $P(\xi_{S\to Q}|G)$ is the probability of seeing $\xi_{S\to Q}$ when the robot targets goal G. The is in line with the notion of action understanding as inverse planning proposed by Baker et al. [15], here $P(\xi_{S\to Q}|G)$ relating to the forward planning problem of finding a trajectory given a goal.

We compute $P(\xi_{S\to Q}|G)$ as the ratio of all trajectories from S to G that pass through $\xi_{S\to Q}$ to *all* trajectories from S to G (Fig.3):

$$P(\xi_{S\to Q}|G) = \frac{\int_{\xi_{Q\to G}} P(\xi_{S\to Q\to G})}{\int_{\xi_{S\to G}} P(\xi_{S\to G})} \tag{5}$$

Following [33], we assume trajectories are separable, i.e. $P(\xi_{X\to Y\to Z}) = P(\xi_{X\to Y})P(\xi_{Y\to Z})$, giving us:

$$P(\xi_{S\to Q}|G) = \frac{P(\xi_{S\to Q}) \int_{\xi_{Q\to G}} P(\xi_{Q\to G})}{\int_{\xi_{S\to G}} P(\xi_{S\to G})} \tag{6}$$

At this point, the robot needs a model of how probable a trajectory ξ is in the eye of an observer. The observer expects the trajectory of minimum cost under C. It is unlikely, however, that they would be completely surprised (i.e. assign 0 probability) by all other trajectories, especially by one ever so slightly different. One way to model this is to make suboptimality w.r.t. C still possible, but exponentially less probable, i.e. $P(\xi) \propto \exp\big(-C(\xi)\big)$, adopting the principle of maximum entropy [33]. With this, (6) becomes:

$$P(\xi_{S\to Q}|G) \propto \frac{\exp\big(-C(\xi_{S\to Q})\big) \int_{\xi_{Q\to G}} \exp\big(-C(\xi_{Q\to G})\big)}{\int_{\xi_{S\to G}} \exp\big(-C(\xi_{S\to G})\big)} \tag{7}$$

Computing the integrals is still challenging. In [37], we derived a solution by approximating the probabilities using Laplace's method (also proposed independently in [38]). If we approximate C as a quadratic, its Hessian is constant and according to Lapace's method, $\int_{\xi_{X\to Y}} \exp\big(-C(\xi_{X\to Y})\big) \approx k\exp\big(-C(\xi^*_{X\to Y})\big)$ (with k a constant and $\xi^*_{X\to Y}$ the optimal trajectory from X to Y w.r.t. C). Plugging this into (7) and using (4) we get:

$$P(G|\xi_{S\to Q}) \propto \frac{\exp\big(-C(\xi_{S\to Q}) - C(\xi^*_{Q\to G})\big)}{\exp\big(-C(\xi^*_{S\to G})\big)} P(G) \tag{8}$$

Much like teleological reasoning suggests [12], this evaluates how efficient (w.r.t. C) going to a goal is through the observed trajectory snippet $\xi_{S\to Q}$ relative to the most efficient (optimal) trajectory, $\xi^*_{S\to G}$. In ambiguous situations like the one in Fig.1, a large portion of $\xi^*_{S\to G}$ is also optimal (or near-optimal) for a different goal, making both goals almost equally likely along it. *This is why legibility does not also optimize C — rather than matching expectation, it manipulates it to convey intent.*

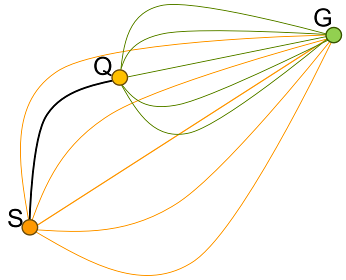


Fig. 3. $\xi_{S\to Q}$ in black, examples of $\xi_{Q\to G}$ in green, and further examples of $\xi_{S\to G}$ in orange. Trajectories more costly w.r.t. C are less probable.

B. *Evaluating and Generating Legibility*

A legible trajectory is one that enables quick and confident predictions. A score for legibility therefore tracks the probability assigned to the actual goal G^* across the trajectory: trajectories are more legible if this probability is higher, with more weight being given to the earlier parts of the trajectory via a function $f(t)$ (e.g. f(t)=T-t, with T the duration of the trajectory):

$$\mathsf{legibility}(\xi) = \frac{\int P(G^*|\xi_{S\to\xi(t)}) f(t)dt}{\int f(t)dt} \tag{9}$$

with $P(G^*|\xi_{S\to\xi(t)})$ computed using C, as in (8).

In situations with multiple possible goals, a robot can make trajectory more and more legible, never reaching a score of 1, and increasing the cost w.r.t. to C more and more. To prevent the robot from going too far away from what the observer expects, we add a regularizer:

$$L(\xi) = \mathsf{legibility}(\xi) - \lambda C(\xi) \tag{10}$$

This brings similar challenges to predictability: knowing the same C, and optimizing a non-convex function (now the maximization of L as opposed to the minimization of C) in high-dimensional spaces.

V. A STUDY OF LEGIBLE AND PREDICTABLE MOTION

The mathematics of predictability and legibility imply that being more legible can mean being less predictable and vice-versa. We set out to verify that this is also true in practice, when we expose subjects to robot motion. We ran an experiment in which we evaluated two trajectories – a theoretically more predictable one ξ_P and a theoretically more legible one ξ_L – in terms of how predictable and legible they are to novices.

A. *Hypothesis*

There exist two trajectories ξ_L and ξ_P for the same task such that ξ_P is more predictable than ξ_L and ξ_L is more legible than ξ_P.

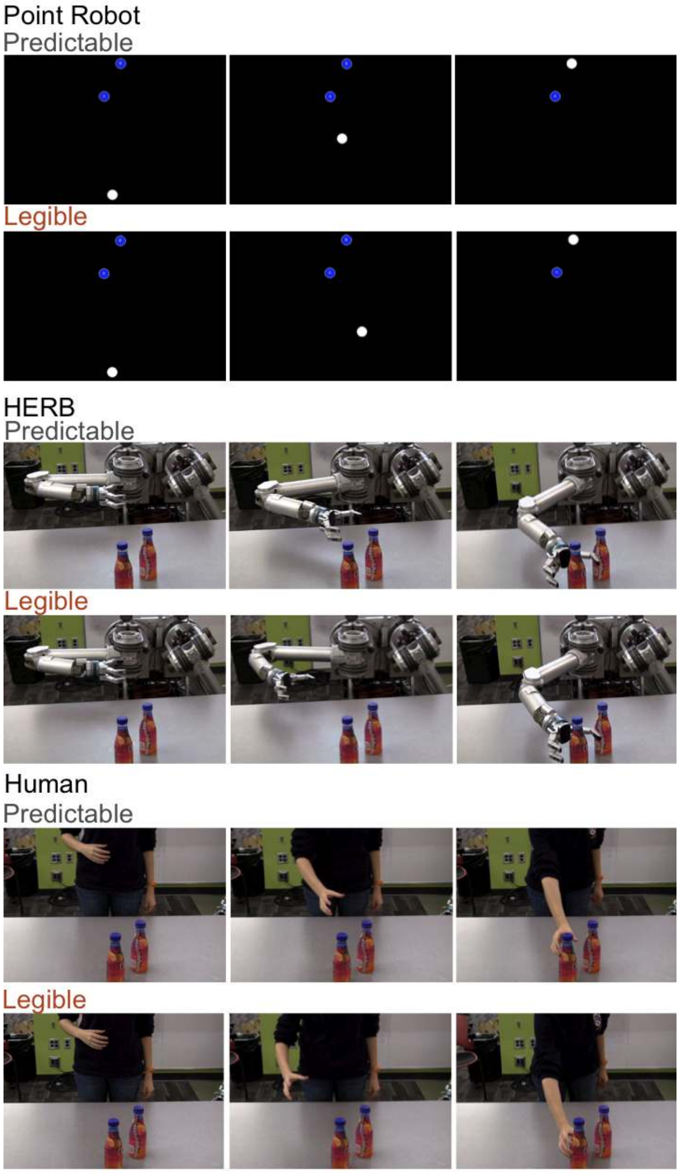


Fig. 4. The trajectories for each character.

B. Experimental Setup

1) Task: We chose a task like the one in Fig.1: reaching for one of two objects present in the scene. The objects were close together in order to make this an ambiguous task, in which we expect a larger difference between predictable and legible motion.

2) Manipulated Variables: **Character:** We chose to use three characters for this task – a simulated point robot, a bimanual mobile manipulator named HERB [16], and a human – because we wanted to explore the difference between humans and robots, and between complex and simple characters.

Trajectory: We hand designed (and recorded videos of) trajectories ξ_P and ξ_L for each of the characters such that predictability$(\xi_P) >$ predictability(ξ_L) according to (2), but legibility$(\xi_P) <$ legibility(ξ_L) according to (9). Verifying for a pair of trajectories that this is true requires assuming a cost function C, and we chose trajectory length (or rather, its quadratic counterpart) in the workspace as a natural representation of efficiency – penalize the robot from taking unnecessarily long paths when the direct one is available. We represent trajectories as vectors of waypoints, and set

$$C_{\text{approx}} = \sum_t ||\xi(t+1) - \xi(t)||^2$$

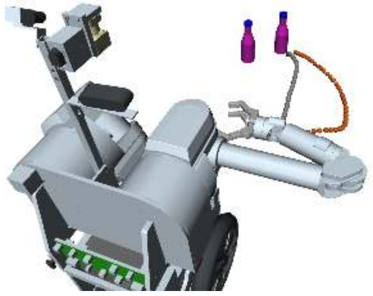


Fig. 5. The end effector trace for the HERB predictable (gray) and legible (orange) trajectories.

While we expect this to be appropriate for the point robot because of its simplicity, we only expect this function to correlate with the real C people expect for the other characters. We describe below several steps we took to eliminate potential confounds arising from this and ensure that the effects we see are actually due to the theoretical difference in the score.

With the HERB character, we controlled for effects of timing, elbow location, hand aperture and finger motion by fixing them across both trajectories. For the orientation of the wrist, we chose to rotate the wrist according to a profile that matches studies on natural human motion [27], [39]), during which the wrist changes angle more quickly in the beginning than it does at the end of the trajectory. Fig.5 plots the end effector trace for the HERB trajectories: the gray one has a larger predictability score $(0.54 > 0.42)$, while the orange one has a higher legibility score $(0.67 > 0.63)$.

With the human character, we used a natural reach for the predictable trajectory, and we used a reach that exaggerates the hand position to the right for the legible trajectory (much like with HERB or the point robot). We cropped the human's head from the videos to control for gaze effects.

3) Dependent Measures: **Predictability:** Predictable trajectories match the observer's expectation. To measure how predictable a trajectory is, we showed subjects the character in the initial configuration and asked them to imagine the trajectory they expect the character will take to reach the goal. We then showed them the video of the trajectory and asked them to rate how much it matched the one they expected, on a 1-7 Likert scale. To ensure that they take the time to envision a trajectory, we also asked them to draw what they imagined on a two-dimensional representation of the scene before they saw the video. We further asked them to draw the trajectory they saw in the video as an additional comparison metric.

Legibility: Legible trajectories enable quick and confident goal prediction. To measure how legible a trajectory is, we showed subjects the video of the trajectory and told them to stop the video as soon as they knew the goal of the character. We recorded the time taken and the prediction.

4) Subject Allocation: We split the experiment into two sub-experiments with different subjects: one about measuring predictability, and the other about measuring legibility.

For the predictability part, the character factor was between-subjects because seeing or even being asked about trajectories for one character can bias the expectation for another. However, the trajectory factor was within-subjects in order to enable relative comparisons on how much each trajectory matched expectation. This lead to three subject groups, one for each character. We counter-balanced the order of the trajectories within a group to avoid ordering effects.

For the legibility part, both factors were between-subjects because the goal was the same (further, right) in all conditions. This leads to six subject groups.

We recruited a total of 432 subjects (distributed approximately evenly between groups) through Amazon's Mechanical Turk, all from the United States and with approval rates higher than 95%. To eliminate users that do not pay attention to the task and provide random answers, we added a control question, e.g. "What was the color of the point robot?" and disregarded the users who gave wrong answers from the data set.

C. Analysis

1) Predictability: In line with our hypothesis, a factorial ANOVA revealed a significant main effect for the trajectory: subjects rated the predictable trajectory ξ_P as matching what they expected better than ξ_L, $F(1,310) = 21.88$, $p < .001$. The main effect of the character was only marginally significant, $F(1,310) = 2,91$, $p = .056$. The interaction effect was significant however, with $F(2,310) = 10.24$, $p < .001$. The post-hoc analysis using Tukey corrections for multiple comparisons revealed, as Fig.6(a) shows, that our hypothesis holds for the point robot (adjusted $p < .001$) and for the human (adjusted $p = 0.28$), but not for HERB.

The trajectories the subjects drew confirm this (Fig.7): while for the point robot and the human the trajectory they expected is, much like the predictable one, a straight line, for HERB the trajectory they expected splits between straight lines and trajectories looking more like the legible one.

For HERB, ξ_L was just as (or even more) predictable than ξ_P. We conducted an exploratory follow-up study with novice subjects from a local pool to help understand this phenomenon. We asked them to describe the trajectory they would expect HERB to take in the same scenario, and asked them to motivate it. Surprisingly, all 5 subjects imagined a different trajectory, motivating it with a different reason.

Two subjects thought HERB's hand would reach from the right side because of the other object: one thought HERB's hand is too big and would knock over the other object, and the other thought the robot would be more careful than a human. This brings up an interesting possible correlation between legibility and obstacle avoidance. However, as Fig.8 shows, a legible trajectory still exaggerates motion away from the other candidate objects even in if it means getting closer to a static obstacle like a counter or a wall.

Another subject expected HERB to not be flexible enough to reach straight towards the goal in a natural way, like a human would, and thought HERB would follow a trajectory made out of two straight line segments joining on a point on the right. She expected HERB to move one joint at a time. We often saw this in the drawn trajectories with the original set of subjects as well (Fig.7, HERB, Expected).

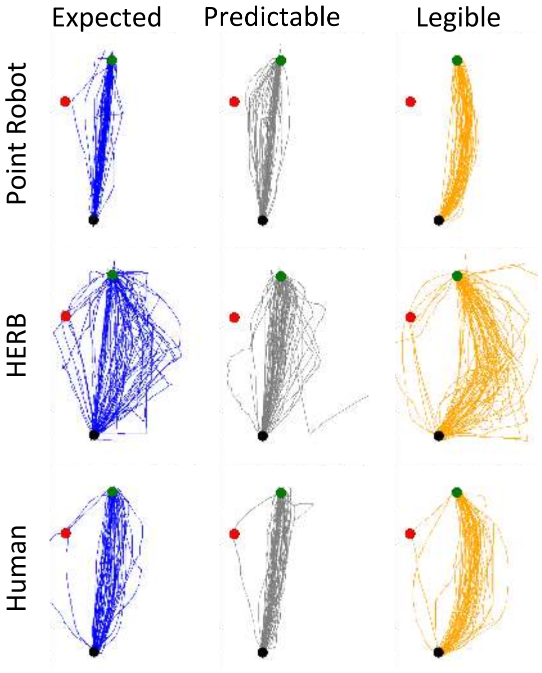


Fig. 7. The drawn trajectories for the expected motion, for ξ_P (predictable), and for ξ_L (legible).

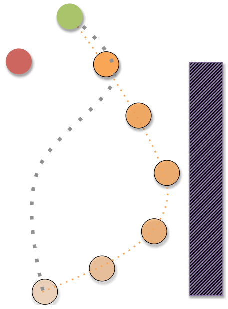


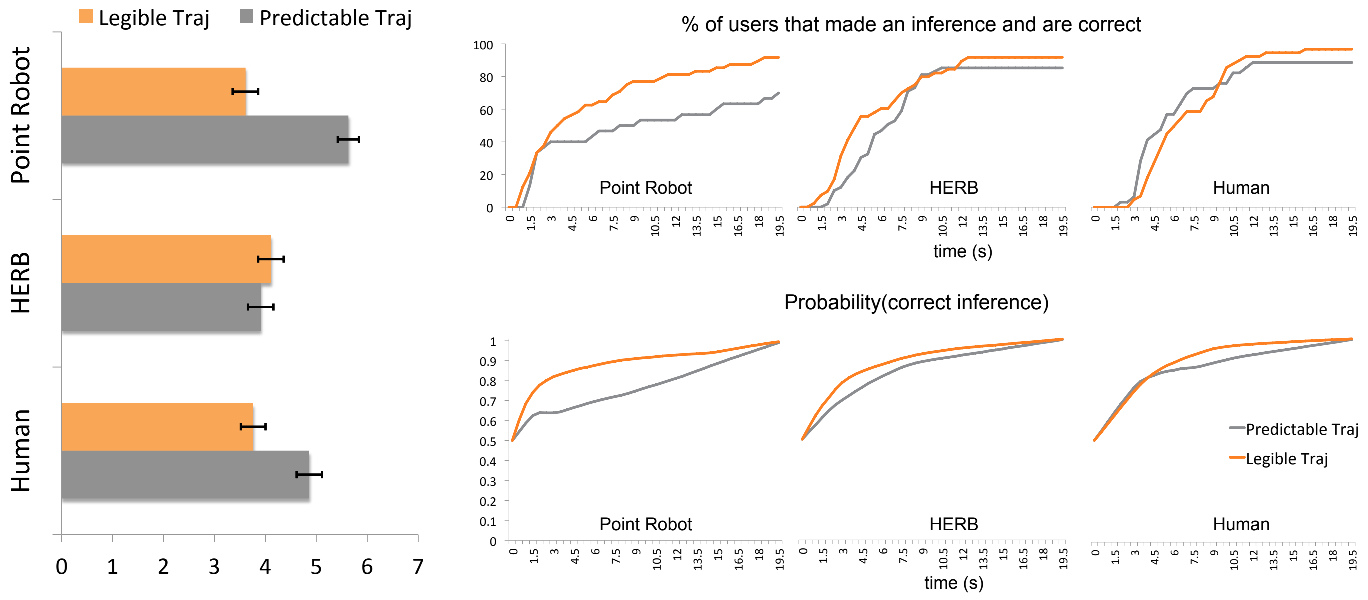
Fig. 8. Legibility is not obstacle avoidance. Here, in the presence of an obstacle that is not a potential goal, the legible trajectory still moves towards the wall, unlike the obstacle-avoiding one (gray trace).

The other subjects came up with interesting strategies: one thought HERB would grasp the bottle from above because that would work better for HERB's hand, while the other thought HERB would use the other object as a prop and push against it in order to grasp the bottle.

Overall, that ξ_P was not more predictable than ξ_L despite what the theory suggested because the cost function we assumed did not correlate to the cost function the subjects actually expected. What is more, every subject expected a different cost function, indicating that a predictable robot would have to adapt to the particulars of a human observer.

2) Legibility: We collected from each subject the time at which they stopped the trajectory and their guess of the goal. Fig.6(b) (above) shows the cumulative percent of the total number of subjects assigned to each condition that made a correct prediction as a function of time along the trajectory. With the legible trajectories, more of the subjects tend to make correct predictions faster.

To compare the trajectories statistically, we unified time and correctness into a typical score inspired by the Guttman



(a) Predictability Rating

(b) Legibility Measures

Fig. 6. (a) Ratings (on Likert 1-7) of how much the trajectory matched the one the subject expected. (b) Cumulative number of users that responded and were correct (above) and the approximate probability of being correct (below).

structure (e.g. [40]): guessing wrong gets a score of 0, and guessing right gets a higher score if it happens earlier.A factorial ANOVA predicting this score revealed, in line with our hypothesis, a significant effect for trajectory: the legible trajectory had a higher score than the predictable one, $F(1, 241) = 5.62, p = .019$. The means were 6.75 and 5.73, much higher than a random baseline of making a guess independent of the trajectory at uniformly distributed time, which would result in a mean of 2.5 – the subjects did not act randomly. No other effect in the model was significant.

Although a standard way to combine timing and correctness information, this score rewards subjects that gave an incorrect answer 0 reward. This is equivalent to assuming that the subject would keep making the incorrect prediction. However, we know this not to be the case. We know that at the end (time T), every subject would know the correct answer. We also know that at time 0, subjects have a probability of 0.5 of guessing correctly. To account for that, we computed an approximate probability of guessing correctly given the trajectory so far as a function of time – see Fig.6(b)(below). Each subject's contribution propagates (linearly) to 0.5 at time 0 and 1 at time T. The result shows that indeed, the probability of making a correct inference is higher for the legible trajectory at all times.

This effect is strong for the point robot and for HERB, and not as strong for the human character. We believe that this might be a consequence of the strong bias humans have about human motion – when a human moves even a little unpredictably, confidence in goal prediction drops. This is justified by the fact that subjects did have high accuracy when they responded, but responded later compared to other conditions. Thus, legible human trajectories would need a stronger emphasis on optimality w.r.t. C (i.e. larger λ in (10)).

VI. Discussion

Limitations: Our work is limited in many ways. Because of the large number of required subjects, our experiment was conducted with videos, instead of exposing subjects to the characters in real life. We made a choice to evaluate legibility by letting users decide when to provide a goal prediction, but it would also be interesting to ask for their prediction and confidence at various points along the trajectory and compare the two. Also, in evaluating predictability, it is possible that when people are asked to make a prediction about how a complex robot like HERB would move, the question itself biases their expectation and they end up building more complex expectations than the obvious, immediate one.

We focused on goal-directed motion only, and did not include other types of motion like emotion expression or gestures, nor other important aspects, like gaze. We also focused our examples as scenes with only two possible goals, and it is true that how legible motion can be is limited in the presence of too many possible goals (as opposed to clutter the observer knows is not a possible goal).

Our experiment was targeted at emphasizing the difference between legibility and predictability in practice, as a confirmation of there difference in theory. Although our results come in support of the models, more experimentation is need in order to attest to their practical utility. Acquiring the cost function C defining user expectation is a still research challenge.

Another limitation and exciting area of future work is that we lack the long-term effects part of the story: since what is predictable can change over time by observing the robot, does legible motion become predictable?

Implications: Collaborative robots must be legible in all collaboration paradigms. They must be legible in shared-workspace collaboration: as the robot reaches for the empty mug, the collaborator should be able to tell early on and reach for the stack of plates instead. They must be legible in robot learning: as the robot is performing the learned task, the teacher should be able to tell early on what the robot will do next and correct the robot when necessary. Finally, they must be legible in assistive teleoperation: as the robot starts using its autonomy to assist in task completion, the operator should be able to tell early on that the robot is correct in what it is doing [37]. Because legibility is fundamentally different (and at times contradictory) from predictability, motion planning for these contexts must switch from a focus on predictability to a focus on legibility, from the cost function that defines what is expected to the one that defines what is intent-expressive.

Acknowledgements

This material is based upon work supported by NSF-IIS-0916557, NSF-EEC-0540865, ONR-YIP 2012, DARPA-BAA-10-28, and the Intel Embedded Computing ISTC. We thank Liz Carter, Katsu Yamane, Andrea Thomaz, Jodi Forlizzi and the members of the Personal Robotics Lab for fruitful discussion and advice.

References

[1] R. Alami, A. Albu-Schaeffer, A. Bicchi, R. Bischoff, R. Chatila, A. D. Luca, A. D. Santis, G. Giralt, J. Guiochet, G. Hirzinger, F. Ingrand, V. Lippiello, R. Mattone, D. Powell, S. Sen, B. Siciliano, G. Tonietti, and L. Villani, "Safe and Dependable Physical Human-Robot Interaction in Anthropic Domains: State of the Art and Challenges," in *IROS Workshop on pHRI*, 2006.

[2] 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*, vol. 28, pp. 21–44, 2010.

[3] M. Tomasello, M. Carptenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: the origins of cultural cognition," *Behavioral and Brain Sciences (in press)*, 2004.

[4] C. Lichtenthäler, T. Lorenz, and A. Kirsch, "Towards a legibility metric: How to measure the perceived value of a robot," in *ICSR Work-In-Progress-Track*, 2011.

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

[6] M. Gielniak and A. Thomaz, "Generating anticipation in robot motion," in *RO-MAN*, 31 2011-aug. 3 2011, pp. 449 –454.

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

[8] 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.

[9] A. Dragan, N. Ratliff, and S. Srinivasa, "Manipulation planning with goal sets using constrained trajectory optimization," in *ICRA*, May 2011.

[10] G. Klien, D. Woods, J. Bradshaw, R. Hoffman, and P. Feltovich, "Ten challenges for making automation a "team player" in joint human-agent activity," *Intelligent Systems*, vol. 19, no. 6, pp. 91 – 95, nov.-dec. 2004.

[11] M. A. Tinker, *Legibility of Print*, 1963.

[12] G. Csibra and G. Gergely, "Obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans," *Acta Psychologica*, vol. 124, no. 1, pp. 60 – 78, 2007.

[13] G. Gergely, Z. Nadasdy, G. Csibra, and S. Biro, "Taking the intentional stance at 12 months of age," *Cognition*, vol. 56, no. 2, pp. 165 – 193, 1995.

[14] G. Csibra and G. Gergely, "The teleological origins of mentalistic action explanations: A developmental hypothesis," *Developmental Science*, vol. 1, pp. 255–259, 1998.

[15] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning appendix," *Cognition*, 2009.

[16] S. Srinivasa, D. Berenson, M. Cakmak, A. Collet, M. Dogar, A. Dragan, R. Knepper, T. Niemueller, K. Strabala, M. V. Weghe, and J. Ziegler, "Herb 2.0: Lessons learned from developing a mobile manipulator for the home," *Proc. of the IEEE, Special Issue on Quality of Life Technology*, 2012.

[17] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1 – 34, 1998.

[18] B. Sodian and C. Thoermer, "Infants' understanding of looking, pointing, and reaching as cues to goal-directed action," *Journal of Cognition and Development*, vol. 5, no. 3, pp. 289–316, 2004.

[19] P. Hauf and W. Prinz, "The understanding of own and others actions during infancy: You-like-me or me-like-you?" *Interaction Studies*, vol. 6, no. 3, pp. 429–445, 2005.

[20] A. T. Phillips and H. M. Wellman, "Infants' understanding of object-directed action," *Cognition*, vol. 98, no. 2, pp. 137 – 155, 2005.

[21] E. J. Carter, J. K. Hodgins, and D. H. Rakison, "Exploring the neural correlates of goal-directed action and intention understanding." *NeuroImage*, vol. 54, no. 2, pp. 1634–1642, 2011.

[22] K. Kamewari, M. Kato, T. Kanda, H. Ishiguro, and K. Hiraki, "Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion," *Cognitive Development*, vol. 20, no. 2, pp. 303 – 320, 2005.

[23] J. Lasseter, "Principles of traditional animation applied to 3d computer animation," in *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987, pp. 35–44.

[24] A. Witkin and M. Kass, "Spacetime constraints," in *Proceedings of the 15th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '88, 1988, pp. 159–168.

[25] M. Gleicher, "Retargeting motion to new characters," in *Proceedings of ACM SIGGRAPH 98*, 1998, pp. 33–42.

[26] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *J Neurosci.*, vol. 5, pp. 1688–1703, July 1985.

[27] L. F and S. JF., "Coordination of arm and wrist motion during a reaching task." *J Neurosci.*, vol. 2, pp. 399–408, April 1982.

[28] B. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469 – 483, 2009.

[29] D. Grollman and O. Jenkins, "Sparse incremental learning for interactive robot control policy estimation," in *ICRA*, 2008.

[30] S. Schaal, A. Ijspeert, and A. Billard, "Computational approaches to motor learning by imitation." *Philos Trans R Soc Lond B Biol Sci*, vol. 358, no. 1431, pp. 537–547, March 2003.

[31] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *ICML*, 2004.

[32] N. Ratliff, J. A. Bagnell, and M. Zinkevich, "Maximum margin planning," in *ICML*, 2006.

[33] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI*, 2008.

[34] N. Ratliff, M. Zucker, J. A. D. Bagnell, and S. Srinivasa, "Chomp: Gradient optimization techniques for efficient motion planning," in *ICRA*, May 2009.

[35] A. Dragan, G. Gordon, and S. Srinivasa, "Learning from experience in manipulation planning: Setting the right goals," in *ISRR*, 2011.

[36] D. Dey, T. Y. Liu, M. Hebert, and J. A. Bagnell, "Contextual sequence prediction with application to control library optimization," in *R:SS*, July 2012.

[37] A. Dragan and S. S. Srinivasa, "Formalizing assistive teleoperation," in *R:SS*, 2012.

[38] S. Levine and V. Koltun, "Continuous inverse optimal control with locally optimal examples," in *ICML '12: Proceedings of the 29th International Conference on Machine Learning*, 2012.

[39] J. Fan, J. He, and S. Tillery, "Control of hand orientation and arm movement during reach and grasp," *Experimental Brain Research*, vol. 171, pp. 283–296, 2006.

[40] G. Bergersen, J. Hannay, D. Sjoberg, T. Dyba, and A. Karahasanovic, "Inferring skill from tests of programming performance: Combining time and quality," in *ESEM*, 2011.