A Model of a Robot's Will Based on Higher-Order Desires

Felix Lindner[1]

Abstract— **Autonomous robots implement decision making capacities on several layers of abstraction. Put in terms of desires, decision making evaluates desires to eventually commit to some most rational one. Drawing on the philosophical literature on volition and agency, this work introduces a conceptual model that enables robots to reason about which desires they want to want to realize, i.e., higher-order desires. As a result, six jointly exhaustive and pairwise disjoint types of choices are defined. Technical evaluation shows how to add a robot's will to its rational decision-making capacity. This guarantees that informed choices are possible even in cases rational decision making alone is indecisive. Further applications to modeling personality traits for human-robot interaction are discussed.**

I. INTRODUCTION

Decision making is undoubtedly one of the most central capacities of autonomous robots. In the taxonomy of robot automation presented by Beer, Fisk, and Rogers [1] the sixth level (out of ten) requires robots to plan and select courses of actions on their own. But also some of the most well-known techniques used to enable robot platforms to exhibit basic capacities such as navigation and locomotion rely on decision making, e.g., path planning, placement selection, collision avoidance etc.

Put in terms of desires, decision making processes enable robots to reason about different and probably competing desires. I will use the term *desire* as a technical term and in a very broad sense throughout this paper: Desires are states of affairs or actions the robot considers more or less worthwhile to achieve or to execute. For example, in the context of AI action planning and BDI-style action selection desires resemble (probably yet uncommited) goals. In the context of placement planning, the candidate placements can count as desires [2]. And also the set of velocities sampled by a collision avoidance algorithm (e.g., [3]) can be understood as a set of desires the robot has to choose from.

More often than not there is not just one unique best choice but competing desires are equally worthwhile. In practice, flipping a coin is the usual approach to tie breaking. Thus, there is one stage of rational decision making followed by a second stage of arbitrariness. While this is reasonable in many cases, it does not display the robot as an autonomous agent, if its commitments are governed (partly) by chance. As a solution, this work proposes a model of a robot's will that disambiguates situations in which multiple desires are equally good (cf., [4]). Figuratively speaking, if a robot wants to realize desire A and B equally strong but it must make a definite choice, it can exercise its will: it wants to want to realize A more than it wants to want to realize B. Consequently, if asked why it sees to realize A rather than B it can state "A and B were equally rational and I wanted A" instead of "A and B were equally rational, therefore I did not care and just took A". It is in this sense that specifications of volition can be expected to enable robots to "exhibit distinctive personality and character" [5, p. 145].

The next section briefly reviews work on voluntarism and on the idea of modeling volition as higher-order desires. Section III presents a formalization which models volition as a strict total order relating first-order desires. An implementation of the model using the answer set programming paradigm is presented in Sect. IV. Sect. V provides concrete examples of application. Technical evaluations conducted on a robot platform demonstrate feasibility of the approach (Sect. VI). Further applications of volition to human-robot interaction are discussed in Sect. VII.

II. VOLITION AND DECISION MAKING

The proposed model of a robot's will—as formalized in the subsequent sections—is inspired by a philosophical discourse on personhood, rational agency, and free will.

In his account on what freedom of action means, Frankfurt [6] introduces the concept of a higher-order desire. First-order desires are very much the same as desires in BDI agents, or (uncommitted) goals, or just options an agent can choose from. Besides being an option, a desire also has a strength, i.e., a desire can be stronger than another. Frankfurt gives the example of a smoker who has the choice between realizing the first-order desire to smoke and realizing the first-order desire to refrain from smoking. Being an addict, the first desire is stronger than the second, in other words, the addict wants to smoke. At the same time the addict can consistently want to want to refrain from smoking, that is, they can have a higher-order desire to refrain from smoking. According to Frankfurt, the addict acts on free will if their higher-order desire determines their final choice.

In Frankfurt's theory, the higher-order desires are strictly ordered. Whereas desires can be equally strong, distinct higher-order desires never stand in conflict to each other [6]. Another property of the higher-order structure is that it emerges from long-term experience with the world. Higher-order desires are rarely questioned in future. In that sense, the volitional structure makes up the person, because it stably separates the desires one identifies with as a person from those that one does not identify with as a person.

One problem such a volition-centered conception of decision making faces is considered by Chang [4]: If volition

[1] Felix Lindner is with the research group on the Foundations of Artificial Intelligence, Department of Computer Science, Albert-Ludwigs-Universität Freiburg, Germany `lindner@informatik.uni-freiburg.de`

were the only source of what an agent should do, then agents would have the power to simply want what they should do. In Chang's account to making choices, she proposes a two-stage decision-making process she calls Hybrid Voluntarism. During the first stage the agent reasons about the rational reasons that speak in favor or against choosing some course of action over another. In cases rational reasoning yields a definitive answer the agent has all-things-considered reasons to stick to that answer. In cases there is no definite answer after careful reasoning, it is the agent's volition that breaks the tie. As Chang [4, p. 266] puts it: "We are slaves to some reasons and masters of others."

Applied to autonomous robots, it can be stated that a robot typically figures out what it has most reason to do rationally. However, to act as a person in the described sense, it needs a structure that represents its volition, i.e., what it wants to want. This structure informs a second stage of decision making to make a choice among equally rational desires. Having said that, volition is distinct from other factors that determine choice, e.g., emotions. Emotions modulate the agent's decision making (e.g., [7], [8]). Volition does not influence the process of rational decision making at all. Instead, it enables robots to reason about how what they rationally want most relates to what they want to want most.

III. A FORMALIZATION OF THE WILL

To implement the concept of will as described in the preceding section, some formalities have to get fixed on a more concrete level. I use first-order logic here to characterize the central concepts and relations. The next section shows that these logical formalizations can be directly implemented and integrated in a robot architecture using answer set programming (e.g., [9]).

A. First-Order Desires

The model of higher-order desires is based on the first-order desires the robot pursues in a decision making situation. $D(x)$ represents that x is a first-order desire. D is here used as a primitive unary predicate. The model assumes that there are always at least two distinct first-order desires. This is no real restriction, because if there is only one desire to do x, there is the possibility to add the desire $y :=$ "to refrain from doing x" as a second desire.

The robot is supposed to have some means to evaluate the desires at hand in order to make a rational decision—for instance, utility functions, cost functions, or decision rules which weigh the pros and cons [10]. No matter what this process might look like in detail the formalization only requires that the robot has the capability to tell the most rational first-order desires from the less rational first-order desires. Let $M(x)$ represent that the first-order desire x is among the most rational first-order desires. Hence, most rational desires are first-order desires (1).

It is well allowed that there are multiple most rational first-order desires but there must be at least one most rational first-order desire (2). Any ordering induced by any objective function will trivially fulfill this requirement even if it assigns the same value to all of the available desires (in this case each of the desires is most rational).

$$\forall x\,[M(x) \rightarrow D(x)] \tag{1}$$

$$\exists x\, M(x) \tag{2}$$

B. Volition as Strictly-Ordered Desires

To model the robot's volition, a binary relation symbol V is introduced: $V(x, y)$ represents that the robot wants to want x more than it wants to want y. To make the notion more clear consider two desires x and y such that $M(x)$ and $\neg M(y)$. In this case it is appropriate to say: "Based on rational reasons, the robot wants to realize x more than it wants to realize y." Conversely, $V(x, y)$ can be put as: "The robot wants to want to realize x more than it wants to want to realize y." Whereas it is possible to want to realize two different desires equally strong it is impossible to want to want to realize two different desires equally strong (cf., [6]). Hence, V is characterized by the axioms (3–6): Only first-order desires are related by V (3), all first-order desires are either related by V or are equal (4), V is asymmetric (5), and V is transitive (6). Thus, V is a strict total order.

$$\forall x, y\,[V(x, y) \rightarrow D(x) \wedge D(y)] \tag{3}$$

$$\forall x, y\,[D(x) \wedge D(y) \rightarrow V(x, y) \vee V(y, x) \vee x = y] \tag{4}$$

$$\forall x, y\,[V(x, y) \rightarrow \neg V(y, x)] \tag{5}$$

$$\forall x, y, z\,[V(x, y) \wedge V(y, z) \rightarrow V(x, z)] \tag{6}$$

C. Varieties of Choices

After the robot has figured out the most rational desires, M, choices can be classified w.r.t. the robot's volition V. $C(x)$ represents that the robot finally made the choice to realize the first-order desire x (thus, choices are first-order desires (7)). The formalization further requires that the robot makes exactly one choice at a time (8).

$$\forall x\,[C(x) \rightarrow D(x)] \tag{7}$$

$$\forall x, y\,[C(x) \wedge C(y) \rightarrow x = y] \tag{8}$$

The first type of choice w.r.t. volition breaks a tie: Suppose there are distinct desires $x, y_1, y_2, \ldots$ such that $M(x) \wedge M(y_1) \wedge M(y_2) \wedge \ldots$, and $V(x, y_1) \wedge V(x, y_2) \wedge \ldots$. That is, the robot considers all of the desires equally rational. Besides that, the robot wants to want to realize x more than it wants to want to realize any of the y_i. If x is the robot's final choice, then I call it *will-determined* (9), otherwise it is named *will-neglecting* (10).

Volition is probably most prominently displayed in situations in which there are more rational reasons to realize y rather than to realize x, and the robot still realizes x because it wants to want to realize x and not y. In such cases I call x *will-enforced* (11). Otherwise, if wanting to realize y is more rational than wanting to realize x, and the robot also wants to want to realize y more than it wants to want to realize x then choosing x is utterly *irrational* (12).

A fifth case exists if realizing desire x is rationally the only best thing to do, and the robot also wants to want to realize x more than anything else. In this case rational decision making automatically matches the will, and one can say then that choosing x is *simply-volitional* (13).

Finally, there might be situations when the robot decides to realize desire x, because realizing x is the only best thing to do, and despite the fact that it wants to want another rationally less desirable desire y more. This case shall be called here *simply-rational* (14).

$$\forall x\, [\mathit{Will\text{-}Determined}(x) \leftrightarrow_{\mathit{def}} C(x) \wedge M(x) \wedge \exists y\, [M(y) \wedge x \neq y] \wedge \neg\exists y\, [M(y) \wedge V(y,x)]] \tag{9}$$

$$\forall x\, [\mathit{Will\text{-}Neglecting}(x) \leftrightarrow_{\mathit{def}} C(x) \wedge M(x) \wedge \exists y\, [M(y) \wedge V(y,x)]] \tag{10}$$

$$\forall x\, [\mathit{Will\text{-}Enforced}(x) \leftrightarrow_{\mathit{def}} C(x) \wedge \neg M(x) \wedge \neg\exists y\, [M(y) \wedge V(y,x)]] \tag{11}$$

$$\forall x\, [\mathit{Irrational}(x) \leftrightarrow_{\mathit{def}} C(x) \wedge \neg M(x) \wedge \exists y\, [M(y) \wedge V(y,x)]] \tag{12}$$

$$\forall x\, [\mathit{Simply\text{-}Volitional}(x) \leftrightarrow_{\mathit{def}} C(x) \wedge M(x) \wedge \neg\exists y\, [M(y) \wedge x \neq y] \wedge \neg\exists y\, V(y,x)] \tag{13}$$

$$\forall x\, [\mathit{Simply\text{-}Rational}(x) \leftrightarrow_{\mathit{def}} C(x) \wedge M(x) \wedge \neg\exists y\, [M(y) \wedge x \neq y] \wedge \exists y\, V(y,x)] \tag{14}$$

D. Some formal Properties of the Formalization

A very desirable theorem of the presented formalization is that the defined volitional attitudes (9)–(14) are jointly exhaustive and pairwise disjoint. Consequently, any possible decision made in any possible situation can unambigously be classified as exactly one of the six cases.

Moreover, in each situation it is possible to make a will-determined, or a simply-volitional, or a simply-rational choice. The reason is this: If there are multiple most rational desires, then, because V is total and strict, one of them must be the most willed one, thus yielding a will-determined choice. If otherwise there is only one most rational desire, then it yields either a simply-volitional or a simply-rational choice. Consequently, if the model is queried to identify, in a concrete decision-making situation, a will-determined, simply-volitional or simply-rational choice, then the result will include exactly one answer. Thus, the robot can always avoid to act irrational, will-enforced, or will-neglecting.

The discussed properties are exploited by the implementation outlined in the next section.

IV. TECHNICAL REALIZATION

The goal is to implement the following two reasoning services based on the proposed formalization of volition:

- *Choice Classification*: Given a choice x the robot has already committed to. The goal is to classify x as falling under one of the six defined volitional attitutes (9)–(14). This is of interest if there is some decision procedure that makes the decisions without taking volition into account, and, subsequently, the robot shall display its volitional attitude towards this decision.
- *Choice Retrieval*: Given one of the six volitional attitutes A from (9)–(14). The goal is to identify decision x of type A. Particularly, using the properties discussed in subsection III-D, this service can be used to identify the will-determined or simply-volitional or simply-rational choice among the available ones. This way, this service serves as a tie-breaking decision procedure.

A. An ASP Encoding of the Volitional Attitutes

To realize both choice classification and choice retrieval, the theory outlined in the preceding section is encoded as an answer set program. Answer set programming has become a major logical programming paradigm. Roughly, the idea is to declaratively describe a problem domain of interest. A solver can then be used to compute all minimal, stable models of the logic program (e.g, see [9]).

As an example for an encoding of the outlined theory, an ASP encoding of definition (9) is shown in Listing 1. Lines 1–3 specify that if i) the chosen desire X is in an answer set, ii) the answer set contains that X is among the most rational desires, iii) there are some most rational desire Y distinct from X in the answer set but iv) none of them is preferred w.r.t. volition, then the fact that X is a will-determined choice must also be included in the answer set. The other definitions (10)–(14) can be encoded accordingly.

```
1  will_determined(X) :- c(X), m(X),
2      #count{Y : m(Y), X != Y} > 0,
3      #count{Y : m(Y), v(Y, X)} = 0.
```

Listing 1. ASP encoding of the theory of will

Clearly, this encoding is an approximation of the first-order formalization. This can for example be seen by the fact that the ASP encoding does not include definitions but rather rules that point into one direction. However, this approximate representation of the theory is sufficient to solve the choice classification and retrieval problems defined above.

B. Encoding of Concrete Decision-Making Situations

Both choice classification and retrieval services take as input the nonempty set of available desires $\mathcal{D} = \{d_1, d_2, \ldots\}$, the nonempty set of the most rational desires $\mathcal{M} = \{d_i, d_j, \ldots\} \subseteq \mathcal{D}$, and the strict volition relation $\mathcal{V} = \{(d_k, d_l), (d_m, d_n), \ldots\} \subseteq \mathcal{D} \times \mathcal{D}$. Additionally, for the choice classification procedure the committed choice d_c must be passed as a fourth argument. For choice retrieval a subset of the six volitional attitutes $\mathcal{A} \subseteq \{\mathit{Will\text{-}Determined}, \mathit{Will\text{-}Neglecting}, \ldots\}$ must be passed as a fourth argument for the procedure to search for all possible choices that are of any of the volitional attitutes stated in $\mathcal{A}$.

Algorithm 1 summarises the main computational steps: First, the given information is encoded as an answer set program. The set of available desires, the most rational desires, and the volition structure are encoded as sets of facts as depicted in Listing 2, lines 1–3. The meaning of these assertions is that every answer set must contain each of these facts. Lines 4 and 5 are to be read as alternatives: For choice classification, a fact like the one in line 4 is asserted. Otherwise, if choice retrieval shall be performed, each of the available desires is a choice candidate. This is encoded

Algorithm 1 Choice Classification and Retrieval

```
function CLASSIFY($\mathcal{D}, \mathcal{M}, \mathcal{V}, d_c$)
    $enc \leftarrow$ ENCODETOCLASSIFY($\mathcal{D}, \mathcal{M}, \mathcal{V}, d_c$)
    $answ \leftarrow$ ASPSOLVER($enc$)
    return EXTRACTVOLITION($answ$)
end function

function RETRIEVE($\mathcal{D}, \mathcal{M}, \mathcal{V}, \mathcal{A}$)
    $enc \leftarrow$ ENCODETORETRIEVE($\mathcal{D}, \mathcal{M}, \mathcal{V}$)
    $answ \leftarrow$ ASPSOLVER($enc$)
    return EXTRACTMATCHINGCHOICES($answ, \mathcal{A}$)
end function
```

in line 5 asserting that exactly one of the desires must be chosen.

```
1 d(d1). d(d2). ... d(dn).
2 m(d1). m(d2). ... m(dj).
3 v(d1, d2). v(d1, d3). ... v(dm, dn).
4 c(d_c).
5 #count{X : c(X) : d(X)} = 1.
```

Listing 2. ASP encoding of the theory of a situation

After the encoding phase, an ASP solver[1] is run. In choice-classification mode, the solver is guaranteed to output exactly one answer set. Because the definitions of the volitional attitutes are jointly exhaustive and pairwise disjoint, this answer set will contain definite information about which of the volitional attitutes the committed choice is of. In the choice-retrieval mode, the solver will output as many answer sets as there are possible choices. Each of these answer sets contains information about what kind of choice it would be if that particular choice was made. This information is finally filtered for those volitional attitutes specified in $\mathcal{A}$.

V. EXAMPLE APPLICATIONS

To demonstrate the application of the theory of volition to robotics, two central components of a mobile robot have been extended by volitional decision making: Reason-based activity placement [2] and sampling-based local path planning using the Dynamic Window Approach (DWA) [3] as implemented in the ROS navigation package[2].

A. Volitional Reason-based Placement Selection

Consider the activity-placement problem instance shown in Fig. 1. The robot's goal is to have its batteries recharged. In this situation there are three candidate placements sp_1, sp_2, sp_3 the robot can use to accomplish its goal. Because using sp_1 would result in blocking a doorway, there is a reason against using sp_1. The decision procedure proposed in [2] that solves the activity-placement problem weighs the pros and cons. In the example situation depicted in Fig. 1, there is a reason against candidate placement sp_1 but no reason against either sp_2 or sp_3. Consequently, the decision procedure will output $sp_1 < sp_2 = sp_3$, i.e., candidate placements sp_2, sp_3 are most rational desires, and sp_1 is not.

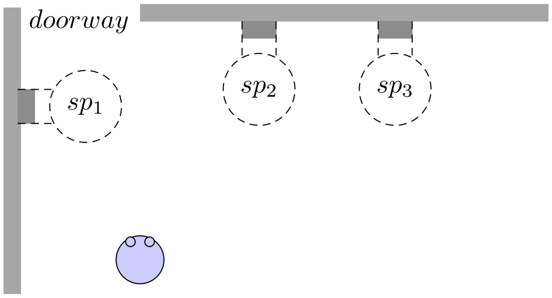

Fig. 1. An instance of an activity-placement problem.

Suppose, for the sake of this example, that according to the robot's volition sp_1 is preferred to sp_2 and to sp_3, and sp_2 is preferred to sp_3. The robot wants to want the leftmost candidate placement more than it wants to want the middle one, and it least wants to want the rightmost placement. This preference for leftmost options is certainly nothing which should be added to the calculation of the rationality or social adequacy of candidate placements. Instead, it can be better modeled as a "personality trait" of the robot using volition.

The encoding of this situation consists of the encoding of definitions (9)–(14) (see Listing 1) together with the information in Listing 3.

```
d(sp1). d(sp2). d(sp3).
m(sp2). m(sp3).
v(sp1, sp2). v(sp1, sp3).
v(sp2, sp3).
```

Listing 3. ASP encoding of the placement-selection situation

As expected, this program has three answer sets: One contains the facts $\{c(sp_2), \textit{will-determined}(sp_2)\}$, a second one contains $\{c(sp_3), \textit{will-neglecting}(sp_3)\}$, and a third one contains $\{c(sp_1), \textit{will-enforced}(sp_1)\}$. Employing the property of the theory discussed in subsection III-D, if the choice retrieval service was asked to retrieve the will-determined or simply-volitional or simply-rational choice, it will return sp_2. Volition breaks the tie, and different to coin flipping, this choice is guaranteed each time the robot faces this decision.

B. Volitional Sampling-based Local Path Planning

After the robot has chosen a placement for recharging, it needs to reach it. To this end, the robot first plans a path to the chosen candidate placement, and then follows the path to the goal. A popular approach to path following while avoiding collisions is the dynamic window approach (DWA) [3]. In each iteration, the DWA samples a set of pairs of translational and rotational velocities, (ν, ω), from the currently reachable velocities. Each sampled (ν, ω) gets then evaluated using an utility function. In the implementation of the DWA in the ROS navigation package, the utility function takes three factors into account: distance to global path, distance to goal, and distance to closest obstacle.

Suppose that the robot is more of the go-getter type of robot. This "personality trait" is constituted by the will

[1] The clingo package was used, see http://potassco.sourceforge.net.

[2] cf., http://wiki.ros.org/dwa_local_planner

to prefer fast transitional velocities over slow ones, and if transitional velocities are equally fast the robot prefers fast rotational velocities over slow ones, and if the rotational velocities are equally fast then it prefers right turns over left turns. Thus, the whole space of possible velocities is strictly ordered with respect to the robot's volition.

Assume that in a concrete iteration the utility function evaluates three sampled velocities as follows: $u((0.5, 0.0)) = 0.1, u((0.1, 0.1)) = 0.1, u((0.1, 0.3)) = 0.5$. Clearly, $(0.1, 0.3)$ is the most rational desire the robot should commit to. For breavity, call $(0.5, 0.0)$ "v_1", $(0.1, 0.1)$ "v_2", and $(0.1, 0.3)$ "v_3". The encoding of this situation consists of the encoding of definitions (9)–(14) (see Listing 1) together with the information in Listing 4.

```
d(v1). d(v2). d(v3).
m(v3).
v(v1, v3). v(v3, v2). v(v1, v2).
```

Listing 4. ASP encoding of the velocity-selection situation

This program has three answer sets: committing to v_3 is simply rational, committing to v_1 is will-enforced, and committing to v_2 is irrational. Employing the property of the theory discussed in Subsect. III-D, if the choice retrieval service was asked to retrieve the will-determined or simply-volitional or simply-rational choice, it will return v_3.

Volition and rational reason do not really interact in this case. The robot should clearly commit to the most rational option, otherwise there would not be any point why to compute utility in the first place. But volition is not absolutely useless here, because the robot could use the fact that its current action is not really what it wants to display its dissatisfaction.

VI. Technical Evaluation

This section investigates rather some of the technical issues that are of interest if the proposed model is used in real robot architectures. The first issue concers the frequency of tie-breaking situations. Clearly, tie breaking is one of the main motivations for modeling volition. But if utility functions were overall decisive and rarely ever assign the same values to multiple samples, this motivation would be of little practical significance. The second issue concerns the computational effort caused by the volition procedure. Local path planning should still be able to operate at a rate of ~ 5 Hz to be reactive in face of occuring obstacles.

A. Experimental Setup

To investigate both frequency of tie breaking and computational footprint, experiments were run on a Turtlebot equipped with a 1.8 Ghz netbook running the dynamic window approach for navigation extended by the volition stage. The robot navigated the same route through an office environment using three different configurations. Each run took approximately one minute driving.

At the time of writing, the default configuration of the DWA as specified in the ROS Turtlebot navigation package sets the number of sampled trajectories per iteration to 120, viz., 20 θ-samples and 6 x-samples. This configuration was taken as a starting point. Moreover, a second configuration was run with 60 sampled trajectories per iteration (12 θ-samples and 5 x-samples), and a third run with 240 sampled trajectories per iteration (24 θ-samples and 10 x-samples).

#Samples	60	120	240
Tie	56 (31%)	91 (55%)	109 (66%)
No Tie	120 (69%)	74 (45%)	56 (33%)
Tie	$M = 40, D = 4$	$M = 48, D = 7$	$M = 59, D = 12$
No Tie	$M = 13, D = 3$	$M = 20, D = 6$	$M = 33, D = 9$

TABLE I
TOP: NUMBER OF TIE AND NO-TIE DECISIONS MADE DURING NAVIGATION. BOTTOM: MEASURED RUNTIMES (MEANS (M) AND MEAN DEVIATIONS (D)) IN MILLISECONDS.

Because tie breaking is the focus here, only will-determined choices are of interest. Therefore, only the most rational desires $\mathcal{M}$ were put into the choice retrieval algorithm. For the same reason, the choice retrieval procedure was only called if there was more than one best trajectory, i.e., a tie. In this case, choice retrieval was invoked by $\textsc{Retrieve}(\mathcal{M}, \mathcal{M}, \mathcal{V}, \{\mathit{will\text{-}determined}\})$, $\mathcal{V} \subset \mathcal{M} \times \mathcal{M}$.

B. Results

The top rows in Table I show the frequency of situations in which the standard DWA did not determine a single rationally best option. As expected, this frequency increases with the number of samples.

The bottom rows in Table I show the runtimes measured during robot navigation. As expected, in tie cases runtimes are higher than in no-tie cases due to the second decision-making stage. Good news is that the mean runtime is well below 200 ms even for twice as many samples as the default configuration. In fact, runtimes in tie cases depend on the number of rationally best samples that constitute the tie, viz., on the size of $\mathcal{M}$. There was one outlier during the 240-samples condition with 82 rationally best trajectories, and it took 220 ms for the volitional stage to make the will-determined choice. The maximum size of $\mathcal{M}$ was 5 in the 60-samples condition, and 9 in the 120-samples condition.

C. Discussion

The results show that tie situations occur frequently. Thus, there is no guarantee that the robot behaves in a coherent way if faced with the same decision several times. Volition ensures that the robot in such cases acts according to its will thereby expressing some personality trait. This comes with extra computational costs, but the DWA algorithm can still keep running at 5 Hz. For applications with tighter time constraints, and to avoid outliers, one could prune situations when the utility function is very indecisive.

For further evaluation, it may also be interesting how the trajectories produced by volitional tie breaking differ from those produced by random tie breaking. However, even in cases they are equal, the volition model still adds that the chosen trajectories are chosen deliberately and thus that the robot has an attitute towards them.

VII. Applications to HRI

So far, the discussed examples highlight the use of volition to model robots as decision makers acting as persons in the sense described in section II. There is significant interest in HRI to model both robot personality [5], [12] and the personality of human interaction partners [13], [14].

Meerbek and colleagues [12] propose a methodology to design a robot's personality. They find that humans naturally connect personality traits to certain robot behaviors. They claim that a well-defined and clearly communicated robot personality helps human interaction partners to build a mental model of the robot. This is in line with the hypothesis that a model of will may lead to more legible robot behavior. Thus, a related research question is how robots can express their personality traits [12], [15].

With respect to modeling human users as persons, Duque and colleagues [13] propose a computational model based on the concept called *persona*. Personas represent specific users' needs along several dimensions like age, proxemic preferences, or user's personality traits. For adaptive human-robot interaction, these representations are mapped to appropriate robot behaviors. Aly and Tapus [14] describe an architecture to match robot personality with human personality. They find, e.g., that extroverts prefer extrovert robots.

Hence, the volition-based model can be applied to several tasks: First, it provides a decision-theoretic semantics to personality traits, e.g., an extroverted robot is modeled by another volition structure than an introverted robot. Second, the same idea can be applied to modeling a human interaction partner, e.g., if a service robot makes rational choices for a human it can use the volition structure representing the human's will to break ties. Third, a robot could choose from different volition structures to match a users' personality. Fourth, as the theory outlined in this paper distinguishes six types of volition-informed choice, a novel question is how these can be displayed, e.g., using a robotic head, and how expressing volition compares to expressing emotions [15].

VIII. Conclusions

This work started out with the observation that robots make decisions permanently and on several layers of abstraction. As the technical experiments demonstrate, rational decision making tends to select multiple options as most rational. This is just fine, because experience tells us that only in the simplest circumstances a unique rationally superior option is available. One branch of philosophy defends that this truth offers humans to exercise their will and thereby display personhood. This article proposes a logics-based model of volition and an integration of volition as part of a two-stage decision procedure. The model enables a robot to reason about how rational reasons compare to its will.

Future research will tackle some limitations of this work. First, the two-stage decision procedure raises the question if volition can be encoded as part of rational decision making to obtain a single-stage tie-free decision procedure. This seems not to be easily done, but this negative claim clearly needs more rigorous analysis. Second, the proposed decision procedure requires that the volition structure is completely defined. A question is how it is possible to define a volition structure on arbitrary desire spaces. One way to think about volition is as emerging from long-term experiences. Thus, processes are needed that learn volition structures based on experience while conserving its formal properties. Particularly, indecisiveness must not pop up on the second stage yielding the need for a third stage etc.

The presented evaluation focus on technical feasibility. In line with the applications outlined in the paper, we currently investigate how a robot can exploit the sixpartite theory of volition to express its attitute towards its choices in a way legible for humans, and hence if the robot is perceived then as acting from will and thus much more like a person.

Acknowledgment

I'd like to thank the reviewers of this paper for their very detailed and thoughtful comments and suggestions.

References

[1] Beer, J. M., Fisk, A. D., Rogers, W. A., Toward a framework for levels of robot autonomy in human-robot interaction, Journal of Human-Robot Interaction, 3(2), 2014, pp. 74–99.

[2] Lindner, F., Eschenbach, C., Affordance-based activity placement in human-robot shared environments, Social Robotics – 5th International Conference, ICSR 2013, Bristol, UK, 2013, pp. 94–103.

[3] Fox, D., Burgard, W., Thrun, S., The dynamic window approach to collision avoidance, IEEE Robotics & Automation Magazine, 4(1), 1997, pp. 23–33.

[4] Chang, R., Voluntarist reasons and the sources of normativity, In Reasons for Action, Cambridge University Press, 2009, pp. 243–271.

[5] Fong, T., Nourbakhsh I., Dautenhahn, K., A survey of socially interactive robots, Robotics and Autonomous Systems 42, 2003, pp. 143–166.

[6] Frankfurt, H. G., Freedom of the will and the concept of a person, The Journal of Philosophy, 68(1), 1971, pp. 5–20.

[7] André, E., Klesen, M., Gebhard, P., Allen, S., Rist, T., Integrating models of personality and emotions into lifelike characters, In Proc. of the workshop on Affect in Interactions—Towards a new Generation of Interfaces in conjunction with the 3rd i3 Annual Conference, Siena, Italy, 1999, pp. 136–149.

[8] Becker-Asano, C., Wachsmuth, I., Affective computing with primary and secondary emotions in a virtual human, Autonomous Agents and Multi-Agent Systems, 20(1), 2010, pp. 32–49.

[9] Eiter, T., Ianni, G., Krennwallner, T., Answer set programming: A primer, In Reasoning Web. Semantic Technologies for Information Systems, 2009, pp. 40–110.

[10] Lindner, F., Soziale Roboter und soziale Räume: Eine Affordanz-basierte Konzeption zum rücksichtsvollen Handeln, PhD Dissertation, University of Hamburg, 2015.

[11] Lindner, F., A conceptual model of personal space for human-aware robot activity placement, Proc. of the 2015 IEEE/RSJ Intern. Conf. on Intelligent Robots and Systems (IROS), 2015, pp. 5770–5775.

[12] Meerbeek, B., Saerbeck, M., Bartneck, C., Iterative design process for robots with personality. In AISB2009 Symposium on New Frontiers in Human-Robot Interaction, 2009, pp. 94–101.

[13] Duque, I., Dautenhahn, K., Koay, K. L., Willcock, L., Christianson, B., A different approach of using personas in human-robot interaction: Integrating personas as computational models to modify robot companions' behaviour. In RO-MAN, 2013, pp. 424–429.

[14] Aly, A., Tapus, A., Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction, Autonomous Robots, 40(2), 2016, pp. 193–209.

[15] Embgen, S., Luber, M., Becker-Asano, C., Ragni, M., Evers, V., Arras, K. O. Robot-specific social cues in emotional body language, In RO-MAN 2012, pp. 1019–1025.