Analyzing Intentional Behavior in Autonomous Agents Under Uncertainty

Filip Cano Córdoba[1], **Samuel Judson**[2], **Timos Antonopoulos**[2], **Katrine Bjørner**[3], **Nicholas Shoemaker**[2], **Scott J. Shapiro**[2], **Ruzica Piskac**[2] and **Bettina Könighofer**[1]

[1]Graz University of Technology

[2]Yale University [3] New York University

{filip.cano, bettina.koenighofer}@iaik.tugraz.at, {samuel.judson, timos.antonopoulos, nick.shoemaker, scott.shapiro, ruzica.piskac}@yale.edu, kbjorner@nyu.edu

Abstract

Principled accountability for autonomous decision-making in uncertain environments requires distinguishing intentional outcomes from negligent designs from actual accidents. We propose analyzing the behavior of autonomous agents through a quantitative measure of the evidence of intentional behavior. We model an uncertain environment as a Markov Decision Process (MDP). For a given scenario, we rely on probabilistic model checking to compute the ability of the agent to influence reaching a certain event. We call this the *scope of agency*. We say that there is evidence of intentional behavior if the scope of agency is high and the decisions of the agent are close to being optimal for reaching the event. Our method applies counterfactual reasoning to automatically generate relevant scenarios that can be analyzed to increase the confidence of our assessment. In a case study, we show how our method can distinguish between 'intentional' and 'accidental' traffic collisions.

1 Introduction

Artificial intelligence (AI)-based autonomous agents play a significant role in diverse facets of society, such as transportation, robotics, medical devices, manufacturing, and more. Ideally, engineers would verify their correctness before deploying them in the real world. However, for various theoretical and practical reasons, formal verification of software for autonomous agents is not often feasible. As a result, autonomous agents might not behave as planned initially and they might cause harm. As we cannot predict when harm will happen, we need to examine the software of the harming autonomous agent *ex post* – after the harm – to assess questions of accountability. While the liability scheme for autonomous agents has yet to be developed, it is plausible to assume that manufacturers of autonomous agents that intentionally harm should be held to a higher standard of accountability than ones that create agents that harm negligently or purely accidentally. Therefore, defining and understanding intention is of paramount significance for establishing accountability. In this paper, we propose a new way of determining whether an autonomous agent has, in fact, acted in a way consistent with the intention to harm.

Historically, symbolic AI produced a large body of work to formally specify and design autonomous agents that were 'rational'. Such agents would explicitly derive decisions based on their beliefs, desires, and intentions (BDI) [Bratman, 1987; Rao and Georgeff, 1995]. Determining whether an autonomous agent has acted with the intention to harm is easy in the case of BDI agents. One just needs to read off the intentions from where they are written in the code. The statistical nature of modern machine-learning-based agents leaves the interpretation of their decision-making in probabilistic settings a far greater challenge, since intentions are not explicitly present in such models.

The traditional view of intention establishes a connection to planning through either cognitive or computational reasoning. Intention is a nuanced legal and philosophical term of art. Here, we use it in the restricted sense of the 'state of the world' the agent plans towards. Whether human or machine, a rational agent with bounded resources must necessarily plan towards a goal to succeed in achieving it [Bratman, 1987; Cohen and Levesque, 1990]. Modern machine-learned agents plan implicitly through techniques like reinforcement learning (RL) [Sutton and Barto, 2018].

This paper considers an autonomous agent operating with other agents within an environment. During the agent's operation, a certain event happened. In the context of holding the agent accountable for such an event, we want to analyze whether the agent acted towards making that event happen.

Problem statement. We concretely model the interactions between the agent and its environment as a Markov Decision Process (MDP). The event under analysis is formalized as a set of states $S_{\mathcal{I}}$ in the MDP. Our goal is to analyze whether the decision-making policy of the agent *shows evidence of intentional behavior towards reaching* $S_{\mathcal{I}}$.

If we assume that the agent has perfect knowledge about the entire world as captured in an MDP, we could simply say: 'There is evidence of intentional behavior towards reaching a state in $S_{\mathcal{I}}$, if the agent implements a policy that maximizes the probability of reaching $S_{\mathcal{I}}$'. However, for any agent acting within a complex environment, this assumption is implausible. For example, the current state information might not be

Find code and experimental details in the accompanying repository https://github.com/filipcano/intentional-autonomous-agents.

precise due to imprecisions in sensor measurements, bounded resources in computing the policy, imprecisions due to abstractions, partial observability, or usage of inaccurate models of other agents. Therefore, we need to consider a certain degree of uncertainty in our assessments.

Method for analyzing intentional behavior. In this paper, we propose a methodology to analyze whether there is evidence that an agent acted intentionally to reach a state in $S_{\mathcal{I}}$. For a given scenario, we use probabilistic model checking to automatically compute the policies that maximize and minimize the probability to reach $S_{\mathcal{I}}$. We use these policies to compute the influence that the agent had to bring about $S_{\mathcal{I}}$. We call this the *scope of agency*. We say that there is evidence of intentional behavior if the scope of agency is high and the decisions of the agent are close to optimal for reaching $S_{\mathcal{I}}$.

To strengthen our evaluation, we make use of a widespread technique in accountability analysis [Wachter *et al.*, 2017], which is analysing a diverse set of relevant *counterfactual scenarios*, and aggregating the evaluation results.

Main Contributions. The main contributions of this paper are the following:

- To the best of our knowledge, we present the first method that analyzes intentional behavior directly from the policies in MDPs.
- We give definitions for *evidence of intentional behavior* in MDPs.
- We propose a method to analyze evidence of intentional behavior of agents in MDPs. Our method uses model checking to automatically relate the agent's policy to any other possible policy. Furthermore, our method applies *counterfactual reasoning* to increase the reliability of the assessment.
- We provide a case study in which we analyze potential intentional behavior in the same scenario for different implementations of driving agents.

2 Preliminaries

Markov Decision Processes. A Markov Decision Process (MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$, where $\mathcal{S}$ is the set of *states*, $\mathcal{A}$ is the set of *actions* and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the *transition function*. A state represents 'one way the world can exist', so any information available to the agent for deciding what to do is included in the state of the MDP. The set $\mathcal{A}$ contains every possible action that can be taken by the agent. The function $\mathcal{P}$ represents the transition to a new environment state that is produced as the result of the agent executing a particular action in a given state.

A *trace* is a finite or infinite sequence of states $\tau = (s_1, s_2, \dots)$. A trace τ is *valid* if for each i, there exists at least one $a_i \in \mathcal{A}$ such that $\mathcal{P}(s_i, a_i, s_{i+1}) > 0$.

The agent is modeled by a memoryless and deterministic *policy* $\pi \colon \mathcal{S} \to \mathcal{A}$ over $\mathcal{M}$ that assigns an action to each state. In Section 7 we discuss how our method can be extended to consider strategies with non-determinism and memory.

Probabilistic Model Checking. Using probabilistic model checking [Clarke *et al.*, 2018], we can compute the exact probability $\mathcal{P}_\pi(\varphi, s)$ of π satisfying a property φ for each state s of the MDP [Kwiatkowska *et al.*, 2011; Hensel *et al.*, 2022]. This property φ will typically be defined in a probabilistic variant of a modal temporal logic, like probabilistic linear temporal logic (PLTL) [Pnueli, 1977].

Let $\Pi \subseteq \{\pi \colon \mathcal{S} \to \mathcal{A}\}$ be a set of policies. We denote the maximum probability of satisfying φ restricted to a policy in Π as $\mathcal{P}_{\max|\Pi}(\varphi, s) = \max_{\pi \in \Pi} \mathcal{P}_\pi(\varphi, s)$. Similarly, we denote the minimum probability as $\mathcal{P}_{\min|\Pi}(\varphi, s)$. In this paper $\varphi := \texttt{Reach}(S)$ encodes the property of *reaching* any state s in a set of states $S \subseteq \mathcal{S}$.

3 Definition of Intentional Behavior in MDPs

In this paper, we assume that we have given a scenario where a certain event happened, like the agent visited a certain location or the agent had a collision with another agent. Our goal is to analyze whether there is evidence the agent intentionally acted towards reaching this event.

In this section, we give the definitions for *evidence of intentional behavior* of policies in the presence of uncertainty. We use an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$ to model the interaction of the agent and the environment. In the following sections, we will then propose and implement a method to analyze intentional behavior according to the definitions of this section.

3.1 Intentions of Agents with Perfect Information

According to [Rao and Georgeff, 1995], an *intention* of an agent is a set of states $S_{\mathcal{I}} \subseteq \mathcal{S}$ the agent committed to reach. Therefore, the agent acts towards reaching $S_{\mathcal{I}}$ to the best of its knowledge.

Let us assume that the agent has perfect knowledge about the environment. For a set of states $S_{\mathcal{I}} \subseteq \mathcal{S}$ to be an intention of an agent, the agent has to implement a policy π that maximizes the probability of reaching $S_{\mathcal{I}}$. Formally, if $S_{\mathcal{I}} \subseteq \mathcal{S}$ is an *intention* of an agent, then $\mathcal{P}_\pi(\texttt{Reach}(S_{\mathcal{I}}), s) = \mathcal{P}_{\max|\Pi}(\texttt{Reach}(S_{\mathcal{I}}), s)$, for any state $s \in \mathcal{S}$.

The policies considered to compute $\mathcal{P}_{\max}$ can be restricted to a set of policies Π, if there are policies that should be excluded for comparison. For example, we may only be interested in policies for comparison that satisfy certain properties like fairness or progress properties.

Definition 1 (Evidence of intentional and non-intentional behavior)**.** *An agent π shows* evidence of intentional behavior *in a state s towards $S_{\mathcal{I}}$ if π maximizes the probability of reaching $S_{\mathcal{I}}$, i.e., $\mathcal{P}_\pi(\texttt{Reach}(S_{\mathcal{I}}), s) = \mathcal{P}_{\max|\Pi}(\texttt{Reach}(S_{\mathcal{I}}), s)$. Otherwise, we say that the agent has* evidence of non-intentional behavior *in state s towards $S_{\mathcal{I}}$.*

3.2 Intentions of Agents Under Uncertainty

The definition of intention given above assumes perfect knowledge about the environment and the agent implementing a policy that is optimal for reaching $S_{\mathcal{I}}$. However, if we want to fully analyze intentional behavior we have to take imprecision and uncertainties into account. Any agent operating in a complex environment needs to make abstractions about

the environmental state and, most likely, only has partial observability. Furthermore, the agent has to make assumptions about the other agents that act within the environment, which may be incorrect. Therefore, we need to relax the definition of intention to take uncertainties into account.

In order to analyze an agent π under uncertainty, we first define the *intention-quotient* $\rho_\pi(s)$ for a state s which represents how close π is to the policy optimal for reaching $S_\mathcal{I}$ from state $s \in \mathcal{S}$.

Definition 2 (Intention-quotient at a state)**.** *For an agent π at a state $s \in \mathcal{S}$, the* intention-quotient *is defined as follows:*

$$\rho_\pi(s) = \frac{\mathcal{P}_\pi(\texttt{Reach}(S_\mathcal{I}), s) - \mathcal{P}_{\min|\Pi}(\texttt{Reach}(S_\mathcal{I}), s)}{\mathcal{P}_{\max|\Pi}(\texttt{Reach}(S_\mathcal{I}), s) - \mathcal{P}_{\min|\Pi}(\texttt{Reach}(S_\mathcal{I}), s)}.$$

In contrast to the case of perfect information, the uncertainty in the agent's knowledge and resources implies uncertainty in the assessment of intentional behavior.

Definition 3 (Evidence of intentional and non-intentional behavior in states)**.** *Given lower and upper thresholds $0 \leq \delta_\rho^L < \delta_\rho^U \leq 1$, we say that there is* evidence of intentional behavior *towards the intention $S_\mathcal{I}$ in the state s, if $\rho_\pi(s) \geq \delta_\rho^U$. Analogously, we say there is* evidence of non-intentional behavior *towards the intention $S_\mathcal{I}$ in the state s, if $\rho_\pi(s) \leq \delta_\rho^L$.*

In case that $\delta_\rho^L < \rho_\pi(s) < \delta_\rho^U$, we say that we have not enough evidence *for intentional behavior.*

By adjusting the thresholds δ_ρ^U and δ_ρ^L, we can control how much discrepancy from the optimal policy under perfect information is allowed in order for π to be still considered as intentional or non-intentional behavior for $S_\mathcal{I}$. In general, the higher the value of the intention-quotient $\rho_\pi(s)$, the more evidence the policy π shows of intentionally trying to reach $S_\mathcal{I}$. The lower the value of $\rho_\pi(s)$, the more evidence the policy π shows on acting without the intention to reach $S_\mathcal{I}$.

An additional source of uncertainty is introduced by the scope of agency of a state. In situations where the agent's actions have little effect on reaching $S_\mathcal{I}$, there is not enough evidence to support a claim of intentional behavior. For this reason, we take the scope of agency into account for our definition of intentional behavior.

Definition 4 (Scope of agency)**.** *The scope of agency $\sigma(s)$ at a state s for intention $S_\mathcal{I}$ is defined as the gap between the best and the worst policy in terms of reaching $S_\mathcal{I}$. Formally, it is given by*

$$\sigma(s) = \mathcal{P}_{\max|\Pi}(\texttt{Reach}(S_\mathcal{I}), s) - \mathcal{P}_{\min|\Pi}(\texttt{Reach}(S_\mathcal{I}), s).$$

The scope of agency of a trace τ *is given by*

$$\sigma(\tau) = \frac{1}{|\tau|} \sum_{s' \in \tau} \sigma(s').$$

If the scope of agency $\sigma(\tau)$ of a trace τ is very low, any assessment about intentional behavior will be very weak.

The above definitions of intentional and non-intentional behavior apply to a single state in the MDP. In order to extend these definitions to traces in the MDP, we aggregate the intention-quotients of the individual states using the scope of agency as the weighting factor.

Definition 5 (Intention-quotient for traces)**.** *For an agent π operating along a trace τ, the intention-quotient $\rho_\pi(S)$ is given as the weighted average*

$$\rho_\pi(\tau) = \frac{1}{\sum_{s \in \tau} \sigma(s)} \sum_{s \in \tau} \sigma(s) \rho_\pi(s).$$

Definition 6 (Evidence of intentional and non-intentional behavior in traces)**.** *Given lower and upper thresholds $0 \leq \delta_\rho^L < \delta_\rho^U \leq 1$, and an agency threshold $0 < \delta_\sigma < 1$, we say that there is* evidence of intentional behavior *towards reaching $S_\mathcal{I}$ along a trace τ, if $\sigma(\tau) \geq \delta_\sigma$ and $\rho_\pi(\tau) \geq \delta_\rho^U$.*

We say that there is evidence of non-intentional behavior *towards reaching $S_\mathcal{I}$ if $\sigma(\tau) \geq \delta_\sigma$ and $\rho_\pi(\tau) \leq \delta_\rho^L$.*

In case that $\delta_\rho^L < \rho_\pi(\tau) < \delta_\rho^U$ or $\sigma(\tau) < \delta_\sigma$, we say that we have not enough evidence *for intentional behavior.*

4 Setting and Problem Statement

In this section, we describe the setting in which we want to analyze intentional behavior and give the problem statement.

Setting. We have a model of the environment in the form of an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$ that captures all relevant dynamics and possible interactions for an agent. We also have a concrete scenario to analyze in the form of a trace $\tau_{ref} = (s_1, \ldots, s_n)$. The trace τ_{ref} is a sequence of visited states in $\mathcal{M}$ that leads to a state in $S_\mathcal{I}$, *i.e.*, $s_n \in S_\mathcal{I}$. The implementation of the agent is given in the form of a policy $\pi\colon \mathcal{S} \to \mathcal{A}$. The underlying intentions of the agent are unknown.

Problem statement. Given this setting, we want to analyze whether there is evidence of the agent acting intentionally, with uncertainty thresholds δ_ρ^D, δ_ρ^U, and δ_σ for the intention-quotient and scope of agency, respectively. Hence, we want to analyze whether there is *evidence of intentional behavior* of the agent π towards intention $S_\mathcal{I}$ in the scenario τ_{ref}.

Example 1. *Let us consider a scenario in which an autonomous car collides with a pedestrian crossing the road. To analyze to which degree the car is accountable for the accident, we are interested in whether causing harm was the intention of the car. In such an example, $\mathcal{M}$ captures all relevant information necessary to analyze the accident, like positions and velocities of car and pedestrian, car dynamics, road conditions, etc. The scenario $\tau_{ref} = (s_1, \ldots, s_n)$ is defined via the sequence of states prior to the collision. The set of states $S_\mathcal{I}$ represents collisions. We want to analyze whether the policy π shows evidence of intentional behavior towards $S_\mathcal{I}$. To avoid unfair comparison with unrealistic policies, we define a set of policies Π that excludes unreasonably slow-moving cars (e.g., cars that stop even though there is no other road user close by).*

5 Methodology

In this section, we propose a concrete methodology to analyze whether there is evidence an agent acted intentionally to reach $S_\mathcal{I}$. Our method is illustrated in Figure 1. As depicted in the figure, we start the *analysis of the given trace τ_{ref}* by computing the intention-quotient $\rho_\pi(\tau_{ref})$ and the scope of agency $\sigma(\tau_{ref})$. If $\sigma(\tau_{ref}) \geq \delta_\sigma$, we can draw conclusions about intentional behavior:

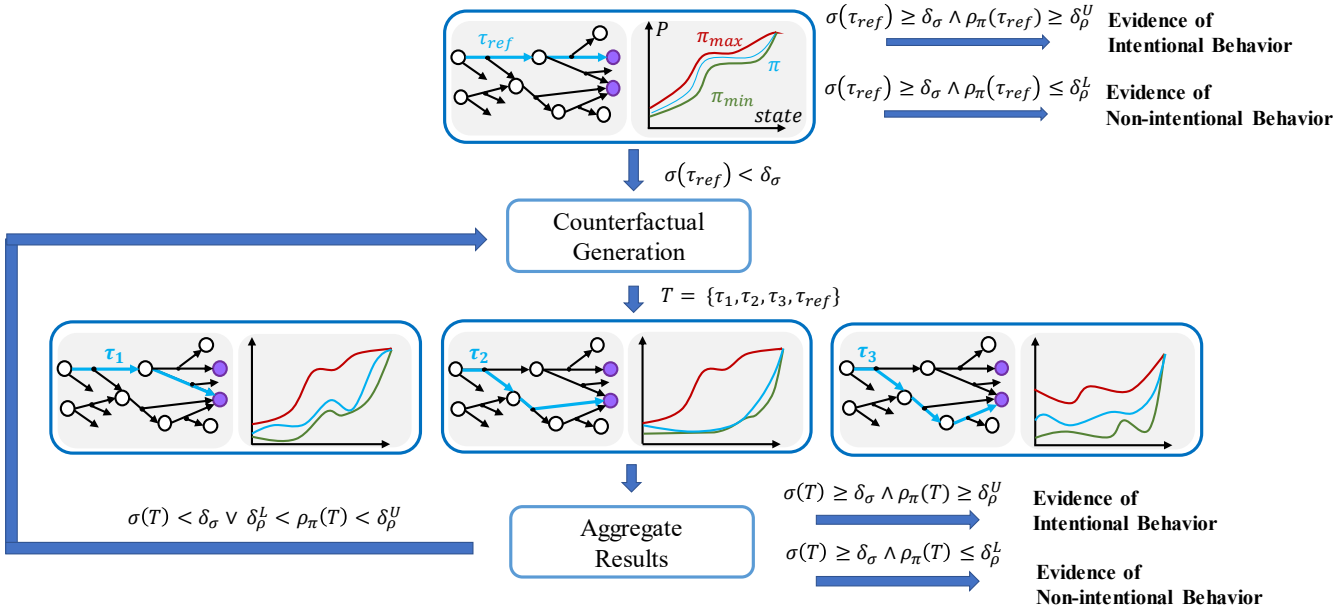

Figure 1: Overview of our approach to analyzing intentional behavior.

- If $\rho_\pi(\tau_{ref}) \geq \delta_\rho^U$, then we conclude that there is evidence of *intentional behavior* towards $S_\mathcal{I}$.
- If $\rho_\pi(\tau_{ref}) \leq \delta_\rho^L$, then we conclude that there is evidence of *non-intentional behavior* towards $S_\mathcal{I}$.

In cases without enough agency, *i.e.*, where $\sigma(\tau_{ref}) < \delta_\sigma$, or where the intention-quotient falls between the lower and upper thresholds, *i.e.*, $\delta_\rho^L < \rho_\pi(\tau_{ref}) < \delta_\rho^U$, we say that we do have not enough evidence to reach a conclusion. In such cases, we propose to *generate* more evidence by analyzing *counterfactual scenarios*.

A counterfactual scenario τ is a scenario close to τ_{ref} according to some distance notion. Our method generates a set of counterfactual scenarios T_{cf} and computes whether there is evidence for intentional or non-intentional behavior for each trace $\tau \in T = T_{cf} \cup \{\tau_{ref}\}$. We fix beforehand the number of counterfactual scenarios to some parameter N.

As before, we draw conclusions about intentional behavior based on the *aggregated results* of the scope of agency $\sigma(T)$ and intention-quotient $\rho_\pi(T)$. If $\sigma(T) < \delta_\sigma$ or $\delta_\rho^L < \rho_\pi(T) < \delta_\rho^U$, there is still not enough evidence for intentional or non-intentional behavior, with $\sigma(T)$ being the scope of agency averaged over all traces in T, and $\rho_\pi(T)$ being the average intention-quotient for the set of traces in T.

In such cases, our algorithm iterates back and extends the set T_{cf} by generating N more counterfactual scenarios to be analyzed. The algorithm stops when enough evidence has been generated to draw a conclusion or when the number of generated counterfactual scenarios exceeds some user-defined limit. In the following, we discuss the generation of counterfactual scenarios in detail.

5.1 Counterfactual Generation

In order to find enough evidence for our assessment of intentional behavior, we generate scenarios that are counterfactuals for τ_{ref}. There are many ways to generate counterfactual traces. We describe here three alternatives, ordered by decreasing the requirement of expert knowledge and involvement.

Counterfactual generation via a human expert. Asking and analyzing counterfactual questions is a standard procedure in accountability processes [Menzies and Beebee, 2020]. Usually, such counterfactual questions are proposed by a domain expert. We transfer this concept to analyzing intentional behavior on MDPs. The counterfactual questions posed by the expert translate to counterfactual traces T_{cf} in $\mathcal{M}$.

Example 2. *Recall Example 1. Some counterfactual questions posed by an expert in the traffic scenario could be: (Q1) What if the car had driven slower? (Q2) What if the pedestrian had been visible earlier? (Q3) What if the road conditions were different? Each of Q1-Q3 translates to a counterfactual trace, which we can analyze in our framework.*

The method of generating counterfactuals using a human expert imposes a heavy burden of work on the expert. Next, we propose two methods to automatically generate counterfactuals to mitigate the need for human effort.

Counterfactual generation using factored MDP. Since $\mathcal{M}$ models the interactions of the agent with its environment, $\mathcal{M}$ is typically given in form of a *factored* MDP. In factored MDPs, the state space of $\mathcal{M}$ is defined in terms of *state variables* $\mathcal{S} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$.

In this approach for counterfactual generation, we assume domain knowledge about which variations of state variables generate interesting counterfactual scenarios. In particular, we assume to know which state variables are *integral state variables* that we want to use in the analysis of intentional behavior, and which variables are *peripheral*. To generate informative counterfactuals, we are interested in changing the

values of the integral state variables.

Example 3. *In Example 1, integral state variables might represent the position and velocity of the car, the position of the pedestrian, the road condition, etc., are integral variables. However, state variables that represent, for example, positions of other pedestrians located behind the car, are most likely labeled as peripheral by a human expert. A counterfactual trace generated from changes in the pedestrians' positions that are not involved in the collision will give no new insights into the assessment of intentional behavior. On the contrary, changing the speed of the car might have a considerable effect on the collision probabilities and may provide an informative counterfactual scenario.*

We automatically generate counterfactual traces by exploring variations of the integral variables. Let the state space be factored as $\mathcal{S} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$, where variables $\mathcal{X}_1, \ldots, \mathcal{X}_k$ are peripheral and $\mathcal{X}_{k+1}, \ldots, X_m$ are integral. For any state $s = (x_1, \ldots, x_m)$, we write its factorization into peripheral and integral variables as $s = (s^{per}||s^{int})$. Let $s^{int}_{ref} = (x_{k+1}, \ldots, x_m)$ be the value of the integral variables at any state of τ_{ref}. We define the set of counterfactual values as:

$$\begin{aligned}\mathrm{Cf}_\epsilon(s^{int}_{ref}) =&\{(y_{k+1}, \ldots, y_m) \in \mathcal{X}_{k+1} \times \cdots \times \mathcal{X}_m \ : \\ &\forall i,\ |x_i - y_i| < \epsilon_i\},\end{aligned}$$

where $\epsilon = (\epsilon_{k+1}, \ldots, \epsilon_m)$ contains for each integral variable the range of variation that is still considered valid. For a given trace $\tau_{ref} = (s_1, \ldots, s_n)$, the counterfactual traces are

$$\begin{aligned}T_C(\tau_{ref}) =&\{(s'_1, \ldots, s'_n) \ : \ \exists s^{int}_{cf} \in \mathrm{Cf}_\epsilon(s^{int}_{ref}),\\ &\forall i = 1 \ldots n : s'_i = (s^{per}_i||s^{int}_{cf}),\\ &(s'_1, \ldots, s'_n) \text{ is valid},\ s'_n \in S_{\mathcal{I}}\}.\end{aligned}$$

Note that the search for counterfactual traces is limited to those integral variables $\mathcal{X}_i$ for which $\epsilon_i > 0$, thus by setting some of the ϵ_i to zero, we can fix their value in the counterfactual generation process.

From T_C, we sample N traces to be used for the counterfactual analysis. For the trace selection, emphasis can be put on traces with higher scopes of agency.

Counterfactual generation using distances on MDPs. This method for generating counterfactual scenarios requires to have given a distance $d\colon \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ defined over states in the MDP. Given such a distance metric d over the states, the set of counterfactual traces is given as

$$\begin{aligned}T_C(\tau_{ref}) =&\{(s'_1, \ldots, s'_n) \ : \ \forall i = 1 \ldots n,\ d(s_i, s'_i) < \eta,\\ &(s'_1, \ldots, s'_n) \text{ is valid}, \quad s_n \in S_{\mathcal{I}}\},\end{aligned}$$

where $\eta > 0$ is a distance that represents states being 'close enough' to be compared as counterfactuals.

In case there is no distance defined in the MDP, there are bisimulation distances that are well defined in any MDP [Song *et al.*, 2016]. They depend on the intrinsic structure of the MDP, defined mainly by similarities in terms of the transition function. The main caveat of this approach is that distances are expensive to compute, and the explanation of why two states are assigned a given distance becomes more obscure to the user.

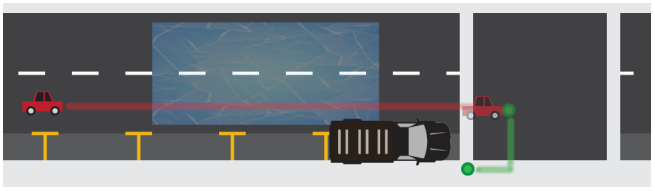

Figure 2: Case-study environment, with scenario τ_{ref} highlighted.

6 Experimental Results

In this section, we showcase our method on a traffic-related scenario related to Examples 1-2, and that is illustrated in Figure 2. In this scenario, a car was driving on a road with a crosswalk. A pedestrian at the crosswalk decided to cross. Close to the crosswalk, there was a parked truck that blocked the visibility of the car. Furthermore, the cold weather conditions made the road slippery, so that braking was less effective than normal. While crossing, the pedestrian was hit by the car. We want to study the behavior of the car for signs of the hit being intentional.

All experiments were executed on an Intel Core i5 CPU with 16GB of RAM running Ubuntu 20.04. We use TEMPEST [Pranger *et al.*, 2021] as our model checking engine.

6.1 Model of Environment

The environment is modeled as an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$. The set of states is a triple $\mathcal{S} = \mathcal{S}^{car} \times \mathcal{S}^{ped} \times \mathcal{S}^{env}$, where $\mathcal{S}^{car}$ models the position and velocity of the car, $\mathcal{S}^{ped}$ models the position of the pedestrian, and $\mathcal{S}^{env}$ models other properties that do not change during a scenario. These properties include the slipperiness factor of the road and the existence of the truck blocking the car's view of the pedestrian.

The car's position is defined via the integers x_c and y_c with $0 \leq x_c \leq 60$ m and $3 \leq y_c \leq 13$ m. The velocity of the car is in $\{0, 1, \ldots, 5\}$ m/s. The position of the car is updated at each step, assuming a uniform motion at the current velocity. The car has the following set of actions $\mathcal{A}$: hitting the brakes, pressing down on the accelerator, and coasting. If the car is on a non-slippery part of the road, accelerating stochastically increases the velocity (by 1 or 2 m/s), braking stochastically decreases the velocity (by 1 or 2 m/s) and coasting maintains or decreases the velocity (by 1 m/s). If the car is on a slippery part of the road, the consequences of the selected action on the velocity change, and include the possibility of no modification to the current velocity for both the actions of braking and accelerating.

The pedestrian's position is given via the integers x_p and y_p with $0 \leq x_p \leq 60$ m and $0 \leq y_p \leq 15$ m. The pedestrian can move 1m in any direction, or not move at all. The probabilities of moving in each direction are given by a stochastic model of the pedestrian, designed in such a way that the pedestrian favors crossing the street through the crosswalk while avoiding being hit by the car. The probabilities in the pedestrian's position update can be influenced by a hesitance factor, which captures how likely it is that the pedestrian puts themselves at a hitting distance from the car. The resulting MDP consists of about 120k states and 400k transitions.

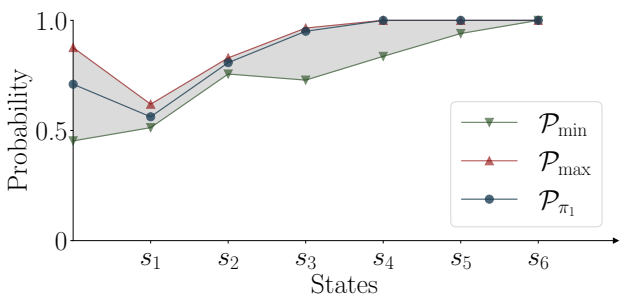

Figure 3: Intention-quotient and scope of agency of τ_{ref}.

	sl_{init}	sl_{end}	sl_{fact}	h_{fact}	vis
Value τ_{ref}	20	45	2.5	0.5	1
Range	$[10, 30]$	$[35, 55]$	$[1, 4]$	$[0.1, 0.9]$	$\{0, 1\}$

Table 1: Ranges to use in counterfactual generation.

$\lvert T\rvert$	6	11	16	21
$\rho_\pi(T)$	0.78 (0.03)	0.81 (0.02)	0.83 (0.02)	0.84 (0.01)
$\sigma_\pi(T)$	0.33 (0.02)	0.44 (0.03)	0.48 (0.01)	0.50 (0.01)
time (s)	53 (16)	147 (42)	227 (32)	318 (64)

Table 2: Results of the counterfactual evaluation.

6.2 Analysis of a Trace

In the described environment, we are given a scenario τ_{ref} as illustrated in Figure 2, and an agent $\pi\colon \mathcal{S} \to \mathcal{A}$. As thresholds to evaluate evidence of intention, we use $\delta_\rho^L = 0.25$, $\delta_\rho^U = 0.75$ and $\delta_\sigma = 0.5$. We restrict the set of policies Π to policies that do not stop the car if no pedestrian is within a range of 15m of the car. The collision states are given by $S_{\mathcal{I}} = \{s \in \mathcal{S} \;:\; |x_p - x_c| \leq 5 \vee |y_p - y_c| \leq 5\}$. Given this setting, we analyze τ_{ref} for evidence of intentional behavior towards reaching $S_{\mathcal{I}}$. Therefore, we first compute the intention-quotient $\rho_\pi(\tau_{ref})$ and the scope of agency $\sigma(\tau_{ref})$.

Results of analysing τ_{ref}. In Figure 3, we give the results of the model checking calls for reaching $S_{\mathcal{I}}$ for states in τ_{ref}. The lower line (▼) represents $\mathcal{P}_{\min}$, the upper (▲) represents $\mathcal{P}_{\max}$ and the line in the middle (●) represents $\mathcal{P}_\pi$ for every state in τ_{ref}. The shaded area, between $\mathcal{P}_{\min}$ and $\mathcal{P}_{\max}$, represents the scope of agency at each state. The figure shows the agent is close to the line of $\mathcal{P}_{\max}$, but the scope of agency is very low, with $\rho_\pi(\tau_{ref}) = 0.73$ and $\sigma(\tau_{ref}) = 0.18$. Since $\sigma(\tau_{ref}) < \delta_\sigma$, our method concludes that there is not enough evidence for intentional behavior yet and moves on to the step of generating counterfactual scenarios.

Counterfactual analysis. We generate counterfactual scenarios by exploiting domain knowledge about integral variables of the MDP. We change the following variables:

- *Slipperiness range*. The street is considered to be slippery between the positions sl_{init} and sl_{end}.
- *Slipperiness factor*. The strength of the slippery effect is measured by the slippery factor sl_{fact}, which is analogous to the inverse of the friction coefficient in classical dynamics. The effect of slipperiness is to make the acceleration and brake less effective, increasing the probability that both acceleration and brake have no effect on the speed of the car. The larger the value of sl_{fact}, the more effect, with $sl_{fact} = 1$ being the minimum value, where the road is considered to be 'not slippery at all'.
- *Hesitancy factor*. The pedestrian, in general, tends to cross the street through the crosswalk. The hesitancy factor modifies the probabilistic model of the pedestrian, to make them more or less prone to put themselves at a hitting distance from the car. A pedestrian with hesitancy factor $h_{fact} = 0$ is a completely cautious pedestrian. On the contrary, a pedestrian with hesitancy factor $h_{fact} = 1$ completely disregards the state of the car.
- *Visibility*. In the given scenario, there is a truck blocking the visibility of the car, corresponding to $vis = 1$. In case $vis = 0$, the visibility block is eliminated.

The variables and the ranges considered for generating counterfactuals are summarized in Table 1.

Results of analyzing counterfactual scenarios. We build the counterfactuals in batches of $N = 5$, by sampling uniformly on the ranges described in Table 1. We show the results in terms of intention-quotient and scope of agency in Table 2. We report the averaged values and standard deviations over 5 runs. As we can see from the table, with 21 traces in T we have $\rho_\pi(T) > \delta_\rho^U = 0.75$ and $\sigma_\pi(T) > \delta_\sigma = 0.5$. Thus, our method concludes that the agent under study does present evidence of intentional behavior to hit the pedestrian.

6.3 Comparative Analysis of Several Agents

In this section, we illustrate how our method can be used to compare different agents in terms of intentional behavior. We compare three different agents π_1, π_2, π_3 in the same scenario τ_{ref}. The agent π_1 corresponds to the policy π in Section 6.2.

In Figure 4 we give the probabilities for reaching $S_{\mathcal{I}}$ for the policies π_1, π_2, π_3 for two different traces: left for τ_{ref}, right for a counterfactual trace $\tau \in T$ with a high scope of agency. The figure illustrates how even a single counterfactual trace can be a powerful tool for distinguishing between policies that seem impossible to differentiate with any confidence in the originally given trace τ_{ref}.

A second insight is illustrated in Table 3. In this table, for each agent π_1, π_2, π_3, we show the number of counterfactuals needed to generate enough evidence of intentional behavior, together with the final values of the intention-quotient and the scope of agency. Both π_1 and π_3 are clear-cut, but for π_2 our algorithm reaches the limit of $|T| = 100$ without finding enough evidence. In this case, the intention-quotient of the agent seems to converge to a value of about 0.53, sitting in the middle of the lower and upper threshold.

	π_1	π_2	π_3
$\lvert T\rvert$	21	100	26
$\rho_\pi(T)$	0.86	0.53	0.14
$\sigma_\pi(T)$	0.52	0.64	0.50

Table 3: Final values of $\rho_\pi(T)$ and $\sigma_\pi(T)$ for different strategies.

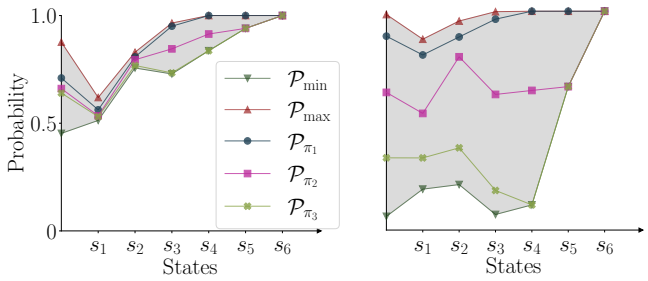

Figure 4: Comparison of τ_{ref} (left) with a high-agency counterfactual scenario (right).

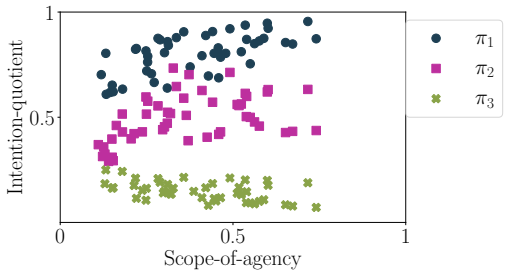

Figure 5: Scatter plot of intention-quotient vs scope of agency for different agents.

Finally, in Figure 5, we show the values of intention-quotient against the scope of agency for 100 counterfactual traces sampled from the ranges in Table 1. This serves as a visual representation of the same facts presented in Table 3, concluding that π_1 (—●—) is clearly showing evidence of intentionally hitting the pedestrian, π_2 (—■—) is showing evidence of intentionally hitting the pedestrian in a lower magnitude, which would be considered enough or not depending on the thresholds, and π_3 (—✖—) is showing clear evidence of acting without the intention of hitting the pedestrian.

7 Discussion

To the best of our knowledge, we present the first method that analyzes intentional behavior directly from the policies given in an MDPs. We believe that our approach has great potential. However, there are aspects that need to be addressed to make the method applicable in challenging scenarios:

- Our method requires having a *correct model of the environment* that captures everything relevant to analyze a scenario. In many cases, such models are not available. Recent work on digital twin technologies [Jones *et al.*, 2020] and the existence of realistic simulators [Dosovitskiy *et al.*, 2017] provides optimism for more and more accurate models of agents and their environment.
- Our method requires the *agent be given as a policy in an MDP*. In case we are given a different implementation, *e.g.*, as a neural network, we would need a sample-efficient method to translate the implementation into a policy in the MDP, at least for the relevant parts of the state space.
- While current probabilistic model checking engines achieve impressive performance [Budde *et al.*, 2021], *computing exact probabilities is costly* (polynomial complexity). An alternative would be to use statistical model checking [Agha and Palmskog, 2018], which is less demanding, albeit also less precise. Statistical model checking has been successfully used to validate autonomous driving modules [Barbier *et al.*, 2019].

General policies. We briefly discuss how to treat policies with memory and non-determinism. Our definitions naturally extend to non-deterministic policies with memory, although it is not evident whether the probabilities required to measure intention-quotients (Definition 2) are easy to compute. Computing extreme probabilities is equally hard for general policies. If the policy has a finite amount μ of memory, $\mathcal{P}_\pi(\texttt{Reach}(S_\mathcal{I}), s)$ can be computed using probabilistic model checking, with a cost of μ times that of the memoryless case [Baier and Katoen, 2008]. In case the non-determinism is unknown to us, to compute $\mathcal{P}_\pi(\texttt{Reach}(S_\mathcal{I}), s)$ we need to sample the decisions of the agent often enough to get an accurate approximation of its decision-making probabilities, making it more costly, although recent heuristics for determinization may help [Ashok *et al.*, 2020].

Knowledge of the agent's beliefs. An intrinsic limitation of studying policies in MDPs is the lack of knowledge of the agent's beliefs about the world. Belief plays a fundamental role in the study of intentions: an agent that intends $S_\mathcal{I}$ must act believing that their acts are a good strategy to reach $S_\mathcal{I}$ [Bratman, 1987]. Belief is also central to the definitions of responsibility and blameworthiness in structural causal models [Chockler and Halpern, 2004; Halpern and Kleiman-Weiner, 2018]. In part for this reason, together with the uncertainties derived from a probabilistic setting, we can only claim incomplete evidence of intentional behavior.

Single-agent setting. In our framework, all relevant parts of the environment are modeled by an MDP, and all the agency in the model is attributed to the agent, *i.e.*, the only actor choosing actions in the MDP is the agent. We argue that this decision is reasonable to study the behavior of an individual agent: from the perspective of an agent, it makes no difference whether the decisions of other actors are governed by a sophisticated policy or by random events in the environment, as long as the MDP model contains accurate transition probabilities. The emergence of intrinsically multi-agent phenomena, like shared intentions in cooperative settings, would require a multi-agent extension of our framework and is left as future work. In particular, we do not explore how to assign moral responsibility to large groups of agents (the so-called "problem of many hands" [Thompson, 1980; Van de Poel, 2015]). Another problem we do not explore is the existence of responsibility voids [Braham and van Hees, 2011], *i.e.*, situations in which a group of agents should be held accountable for an outcome, while at the same time, no individual agent intended that outcome.

8 Related Work

Intention in artificial intelligence. We borrow the concept of intention as a set of states to reach from standard BDI literature [Rao and Georgeff, 1991; Rao and Georgeff, 1995]. Closest related to our work is [Simari and Parsons, 2011], where the authors develop a mapping from the BDI formalism to the MDP formalism. The mapping they propose on intentions to policies in MDPs yields a definition of intentions in MDP similar to our Definition 1 of intentional behavior under perfect knowledge. In contrast to our work, Simari and Parsons focus on optimal policies in MDPs and their correspondence to plans following a certain intention in the BDI model. Therefore, their mapping holds only for optimal policies and cannot be applied to agents with suboptimal policies.

A central element in the definition of intention is commitment: an agent should not reconsider its intentions too often [Cohen and Levesque, 1990]. Although we do not model reconsideration as it relates to time, the intention quotient ρ_π can be interpreted as a quantitative measure of the agent's commitment to reach a certain state.

Responsibility and accountability. The concept of intention of rational agents, both humans and non-humans, has been the subject of extensive study in the context of philosophy of action [Anscombe, 1957; Mele, 1992; Bratman, 1999] as well as in its relation to moral responsibility [Braham and van Hees, 2012; Scanlon, 2010]. The concept of agency is a necessary element in assigning responsibility, leading to issues when the agency is diluted among many individuals [Shapiro, 2014; Braham and van Hees, 2011]. There is an ongoing debate in the philosophy of mind, between those that consider that an agent's reasoning is sufficient to explain their actions [Quine, 1969], and those who maintain that extrinsic information must be imported through a "Principle of Charity" [Davidson, 1963]. By building a model of the agent's knowledge (the MDP) to inquire about their behavior, we are assuming the latter position. Recent work attempts to answer similar questions from the former [Judson *et al.*, 2023].

Causality and blame attribution. A basic element for a complete accountability process is the study of *causality* [Halpern and Pearl, 2005a; Halpern and Pearl, 2005b]. The foundational work of [Chockler and Halpern, 2004] introduced a quantitative notion of causality, by studying degrees of responsibility and blame. Responsibility and blame allocation has been extensively developed in the context of non-probabilistic structures (see, e.g., [Aleksandrowicz *et al.*, 2017] for the characterization of complexity or [Yazdanpanah and Dastani, 2016] for a multi-agent framework). More recent and more closely related to our approach is the work of [Baier *et al.*, 2021], studying responsibility and blame in Markov models. The study of harm from a causality perspective is also gaining attention recently, with [Beckers *et al.*, 2022] studying harm from an actual causality perspective, and [Richens *et al.*, 2022] studying harm from a probabilistic perspective, heavily relying on counterfactuals. Counterfactual analysis [Lewis, 2013] is a key concept in causality [Pearl, 2009], used in an analogous way as our generation of counterfactual scenarios. We go one step further by relating the implementation of the agent to the best and worst implementation for reaching an intended event. Another recent approach to blame attribution is [Triantafyllou *et al.*, 2021], which studies multi-agent Markov decision processes from a game-theoretic perspective.

Policy-discovery methods. Since the popularization of reinforcement learning, there exist several methods for obtaining representations of a black-box agent, by studying traces of such agents. In inverse reinforcement learning [Ng and Russell, 2000; Agha and Palmskog, 2018], the agent is assumed to be maximizing an unknown reward function, and the objective is to find the reward function that best explains the agent's performance over a set of traces. These methods could potentially be used as a pre-processing step to apply our framework to black box agents. In any case, the obtained representations must be accurate enough before using them for any accountability process.

Explainability. One of the most influential works in *explainability* of AI is [Miller, 2019], which studies how explainability should rely on concepts from social sciences. More recently [Winikoff *et al.*, 2021] uses the built-in notions of desire, beliefs and intentions to study explainability of BDI models, relying on concepts from the sociology literature. While the main paradigm in explainable reinforcement learning is applying techniques from explainable machine learning [Puiutta and Veith, 2020], our analysis of intentional behavior can be used as a method to aid the interpretability of agents operating in MDPs, using concepts from the philosophy of action [Bratman, 1987].

9 Conclusion & Future Work

In this paper, we analyzed policies in MDPs with respect to intentional behavior taking uncertainties into account. Our method uses probabilistic model checking to automatically compute the best and worst possible policy for reaching a set of intended states. We assess evidence of intentional behavior in a policy by relating it to the best and worst policies, and use counterfactual analysis to generate more evidence if needed.

In future work, we want to extend our current analysis by considering a multitude of possibly conflicting intentions of the agent. Another interesting line of work is to extend the study of intentional behavior to multi-agent systems, in which cooperative or competitive intentions may arise. We also want to study long executions, where the agent has time for reconsideration. Furthermore, we want to implement our framework to study reinforcement learning agents in challenging application areas.

Acknowledgements

This work was supported in part from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 956123 - FOCETA, by the State Government of Styria, Austria – Department Zukunftsfonds Steiermark, the Office of Naval Research (ONR) of the United States Department of Defense through an National Defense Science and Engineering Graduate (NDSEG) Fellowship, and by the National Science Foundation (NSF) awards CCF-2131476 , CCF-2106845, and CCF-2219995. We also thank Lukas Posch for his help in setting up TEMPEST.

References

[Agha and Palmskog, 2018] Gul Agha and Karl Palmskog. A survey of statistical model checking. *ACM Transactions on Modeling and Computer Simulation*, 28(1):1–39, 2018.

[Aleksandrowicz *et al.*, 2017] Gadi Aleksandrowicz, Hana Chockler, Joseph Y. Halpern, and Alexander Ivrii. The computational complexity of structure-based causality. *Journal of Artificial Intelligence Research*, 58:431–451, 2017.

[Anscombe, 1957] Gertrude Elizabeth Margaret Anscombe. *Intention*. Harvard University Press, 1957.

[Ashok *et al.*, 2020] Pranav Ashok, Mathias Jackermeier, Pushpak Jagtap, Jan Křetínský, Maximilian Weininger, and Majid Zamani. dtControl: decision tree learning algorithms for controller representation. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control (HSCC'20)*, pages 17:1–17:7. ACM, 2020.

[Baier and Katoen, 2008] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*, chapter 10. Probabilistic Systems. MIT Press, 2008.

[Baier *et al.*, 2021] Christel Baier, Florian Funke, and Rupak Majumdar. Responsibility Attribution in Parameterized Markovian Models. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*, pages 11734–11743. AAAI Press, 2021.

[Barbier *et al.*, 2019] Mathieu Barbier, Alessandro Renzaglia, Jean Quilbeuf, Lukas Rummelhard, Anshul Paigwar, Christian Laugier, Axel Legay, Javier Ibañez-Guzmán, and Olivier Simonin. Validation of Perception and Decision-Making Systems for Autonomous Driving via Statistical Model Checking. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV'19)*, pages 252–259. IEEE, 2019.

[Beckers *et al.*, 2022] Sander Beckers, Hana Chockler, and Joseph Y Halpern. A Causal Analysis of Harm. In *Advances in Neural Information Processing Systems (NeurIPS'22)*, volume 35, pages 2365–2376. Curran Associates, Inc., 2022.

[Braham and van Hees, 2011] Matthew Braham and Martin van Hees. Responsibility voids. *The Philosophical Quarterly*, 61(242):6–15, 2011.

[Braham and van Hees, 2012] Matthew Braham and Martin van Hees. An Anatomy of Moral Responsibility. *Mind*, 121(483):601–634, 2012.

[Bratman, 1987] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

[Bratman, 1999] Michael E. Bratman. *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge Studies in Philosophy. Cambridge University Press, 1999.

[Budde *et al.*, 2021] Carlos E. Budde, Arnd Hartmanns, Michaela Klauck, Jan Křetínskỳ, David Parker, Tim Quatmann, Andrea Turrini, and Zhen Zhang. On Correctness, Precision, and Performance in Quantitative Verification. In *Proceedings of the 9th International Symposium on Leveraging Applications of Formal Methods (ISoLA'21)*, pages 216–241. Springer, 2021.

[Chockler and Halpern, 2004] Hana Chockler and Joseph Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[Clarke *et al.*, 2018] Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith, and Roderick Bloem. *Handbook of Model Checking*. Springer, 2018.

[Cohen and Levesque, 1990] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2):213–261, 1990.

[Davidson, 1963] Donald Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.

[Dosovitskiy *et al.*, 2017] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning (CoRL'17)*, volume 78, pages 1–16. PMLR, 2017.

[Halpern and Kleiman-Weiner, 2018] Joseph Y. Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pages 1853–1860. AAAI Press, 2018.

[Halpern and Pearl, 2005a] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

[Halpern and Pearl, 2005b] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

[Hensel *et al.*, 2022] Christian Hensel, Sebastian Junges, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. The probabilistic model checker storm. *International Journal on Software Tools for Technology Transfer*, 24(4):589–610, 2022.

[Jones *et al.*, 2020] David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020.

[Judson *et al.*, 2023] Samuel Judson, Matthew Elacqua, Filip Cano Córdoba, Timos Antonopoulos, Bettina Könighofer, Scott J. Shapiro, and Ruzica Piskac. 'Put the Car on the Stand': SMT-based Oracles for Investigating Decisions. *preprint, arXiv:2305.05731*, 2023.

[Kwiatkowska *et al.*, 2011] Marta Kwiatkowska, Gethin Norman, and David Parker. PRISM 4.0: Verification of Probabilistic Real-time Systems. In *Proceedings of the 23rd International Conference on Computer Aided Verification (CAV'11)*, pages 585–591, 2011.

[Lewis, 2013] David Lewis. *Counterfactuals*. John Wiley & Sons, 2013. Originally published in 1973.

[Mele, 1992] Alfred R. Mele. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press, 1992.

[Menzies and Beebee, 2020] Peter Menzies and Helen Beebee. Counterfactual Theories of Causation. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2020.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Ng and Russell, 2000] Andrew Y. Ng and Stuart Russell. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML'00)*, pages 663–670. Morgan Kaufmann, 2000.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[Pnueli, 1977] Amir Pnueli. The temporal logic of programs. In *Proceedings of the 18th Symposium on Foundations of Computer Science (FOCS'77)*, pages 46–57. IEEE, 1977.

[Pranger *et al.*, 2021] Stefan Pranger, Bettina Könighofer, Lukas Posch, and Roderick Bloem. TEMPEST - Synthesis Tool for Reactive Systems and Shields in Probabilistic Environments. In *Proceedings of the 19th International Symposium on Automated Technology for Verification and Analysis (ATVA'21)*, volume 12971 of *Lecture Notes in Computer Science*, pages 222–228. Springer, 2021.

[Puiutta and Veith, 2020] Erika Puiutta and Eric Veith. Explainable Reinforcement Learning: A Survey. In *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE'20)*, pages 77–95. Springer, 2020.

[Quine, 1969] Willard Van Orman Quine. *Ontological Relativity and Other Essays*. Columbia University Press, 1969.

[Rao and Georgeff, 1991] Anand S. Rao and Michael P. Georgeff. Modeling Rational Agents within a BDI-Architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann, 1991.

[Rao and Georgeff, 1995] Anand S. Rao and Michael P. Georgeff. BDI Agents: From Theory to Practice. In *Proceedings of the 1st International Conference on Multiagent Systems (ICMAS'95)*, pages 312–319. MIT Press, 1995.

[Richens *et al.*, 2022] Jonathan Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm. In *Advances in Neural Information Processing Systems (NeurIPS'22)*, volume 35, pages 36350–36365, 2022.

[Scanlon, 2010] Thomas Michael Scanlon. *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press, 2010.

[Shapiro, 2014] Scott J. Shapiro. Massively Shared Agency. *Rational and Social Agency: The Philosophy of Michael Bratman*, pages 257–293, 2014.

[Simari and Parsons, 2011] Gerardo I. Simari and Simon D. Parsons. *Markov Decision Processes and the Belief-Desire-Intention Model: Bridging the Gap for Autonomous Agents*. Springer Science & Business Media, 2011.

[Song *et al.*, 2016] Jinhua Song, Yang Gao, Hao Wang, and Bo An. Measuring the Distance between Finite Markov Decision Processes. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems (AAMAS'16)*, pages 468–476. ACM, 2016.

[Sutton and Barto, 2018] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[Thompson, 1980] Dennis F Thompson. The Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review*, 74(4):905–916, 1980.

[Triantafyllou *et al.*, 2021] Stelios Triantafyllou, Adish Singla, and Goran Radanovic. On Blame Attribution for Accountable Multi-Agent Sequential Decision Making. In *Advances in Neural Information Processing Systems (NeurIPS '21)*, volume 34, pages 15774–15786. Curran Associates, Inc., 2021.

[Van de Poel, 2015] Ibo Van de Poel. The problem of many hands. In *Moral responsibility and the problem of many hands*, pages 50–92. Routledge, 2015.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2017.

[Winikoff *et al.*, 2021] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. Why bad coffee? explaining BDI agent behaviour with valuings. *Artificial Intelligence*, 300:103554, 2021.

[Yazdanpanah and Dastani, 2016] Vahid Yazdanpanah and Mehdi Dastani. Distant group responsibility in multi-agent systems. In *Proceedings of the International Conference on Principles of Practice in Multi-Agent Systems (PRIMA'16)*, pages 261–278. Springer, 2016.