Research Article

Thomas Hellström*

The relevance of causation in robotics: A review, categorization, and analysis

https://doi.org/10.1515/pjbr-2021-0017
received December 15, 2020; accepted March 23, 2021

Abstract: In this article, we investigate the role of causal reasoning in robotics research. Inspired by a categorization of human causal cognition, we propose a categorization of robot causal cognition. For each category, we identify related earlier work in robotics and also connect to research in other sciences. While the proposed categories mainly cover the sense–plan–act level of robotics, we also identify a number of higher-level aspects and areas of robotics research where causation plays an important role, for example, understandability, machine ethics, and robotics research methodology. Overall, we conclude that causation underlies several problem formulations in robotics, but it is still surprisingly absent in published research, in particular when it comes to explicit mentioning and using of causal concepts and terms. We discuss the reasons for, and consequences of, this and hope that this article clarifies the broad and deep connections between causal reasoning and robotics and also by pointing at the close connections to other research areas. At best, this will also contribute to a "causal revolution" in robotics.

Keywords: robotics, causal cognition, causal reasoning, causal inference, causation, causality, counterfactual, intelligence

1 Introduction

Causal reasoning is generally considered to be a crucial human cognitive competence necessary for a range of tasks such as predictions, diagnoses, categorization, action planning, decision-making, and problem solving [1]. The development of causal understanding has been described as a major part of human evolution [2,3], and causal functionality has also occasionally been recognized as a necessity for the development of intelligent robots [4].

The purpose of this article is to provide a structured analysis of how the broad area of causality relates to the equally broad area of robotics. Inspired by a categorization of human causal cognition [5], we propose a categorization of robot causal cognition. We also identify several additional aspects of robotics research where causation plays, or may play, an important role. We review related earlier work in robotics and also identify connections to earlier work on causation in other sciences.

Overall, we find very little earlier work in robotics where causation is explicitly mentioned. However, we identify several interesting connections between large sub-areas of robotics and causal concepts. For example, task planning and several common learning paradigms rely on, often implicit, causal assumptions and principles.

The remainder of this article is organized as follows: Section 2 provides an overview of how causation is, and has been, dealt with in other sciences. In Section 3, we analyze the role of causal reasoning in robotics together with links to earlier work. The reasons for, and consequences of, the relatively weak appearance of causation in published robotics research is finally discussed in Section 4.

Basic terminology often varies between different research fields and even between authors in the same field. In this article, we interchangeably use the terms *Causal reasoning* and *Causal cognition* (as often used in psychology [1,5,6]) to denote the union of two processes of (a) learning models of causal relations and (b) using and drawing conclusions from causal models. Process (a) is denoted here as *Causal learning*, and process (b) is denoted as *Causal inference* (in computer science and statistics, the latter term is often used to denote the union of (a) and (b) [7–9]).

The presented work is applicable to all kinds of robots but is particularly relevant for robots that interact closely with humans, other robots, or the environment.

* **Corresponding author: Thomas Hellström,** Department of Computing Science, Umeå University, Umeå, Sweden, e-mail: thomas.hellstrom@umu.se

2 Causation outside of robotics

Causation has for a very long time been an active topic in several sciences, and we will in this section summarize how causation is, and has been, dealt with in philosophy, statistics, computer science, cognitive psychology, and machine learning. Each field has on its own produced huge amounts of research, and our summary is, for that reason, very selective – focusing on the concepts and notions that are relevant for causation in conjunction with robotics.

2.1 Philosophy

Causation is one of these concepts that continue to puzzle us, even after millennia of considerable intellectual effort. Viewpoints among scholars still vary widely, and there is not even an agreement that causation exists, let alone on what it is. Nevertheless, people use causal expressions daily, with a reasonably agreed upon meaning. Most people are also quite clear about the difference between causation and correlation, even if they not necessarily use these terms. As an example, let us imagine that the owner of an ice-cream shop observes the relation between electricity consumption and ice-cream sales during a hot summer month. It turns out that sales goes up during days with high electricity consumption and *vise versa*. Does that mean that the owner should switch on all lights in the shop in order to increase sales? Most of us would probably agree that this would be a bad idea and that the reason for the observed correlation between ice-cream and electricity consumption is a third variable, the outdoor temperature, which makes people buy more ice-cream, but also increases the electricity consumption, since the A/C has to work harder during hot days.

Not surprisingly, philosophers have not settled with such intuitive notions of causation. In the remainder of this subsection, we will merely scratch the surface of the substantial body of work on the nature of causation, focusing on philosophical theories that are relevant for the continued analysis and discussion.

Woodward [10] identifies two main kinds of philosophical views on what causation is. The *difference-making* views focus on how causes affect what happens in the world, such that the state of the world differs depending on whether the cause occurs or not. For example, a robot vacuum cleaner starting up in the middle of the night might be regarded as the cause of the owner waking up. The other kind of views are denoted as the *geometrical-mechanical theories* and focus on the idea that causes somehow are connected to their effects, often through transmission of a "force" from cause to effect. A typical example is a billiard ball hitting another ball, thereby stopping while the other ball starts moving due to the "transfer" of energy or momentum.

As is most often the case with theories on causation, compelling counterexamples for both views are easily formulated. The billiard ball example can be easily questioned by referring to fundamental physics principles of relativity of motion. One ball moving toward a second is equivalent to the second moving toward the first, but according to the geometrical–mechanical view different balls will be assigned as the cause in the two cases. A classical example, challenging the difference-making view, is a firing squad, where ten soldiers shoot a victim who dies. However, for each one of the soldiers can be said that he or she did not cause the death, since the outcome would have been the same had the soldier not fired the gun. Each individual soldier did not make a difference. On the other hand, there are lots of examples of events that make a difference, but are not commonly regarded as causes. Big bang, certainly made several differences, but it is usually not referred as a cause to things happening now, billions of years later. A more recent example is given in ref. [11, p. 127], where is it noted that having a liver is a necessary, albeit not sufficient, condition for having cirrhosis. Therefor, having a liver is a difference-maker for having that disease, even if is does not make sense to say that having a liver caused the disease.

This approach of reasoning about necessary and sufficient conditions was refined by Mackie [12,13], who argued that what we normally mean when we say that something is a cause is that it is "an Insufficient but Non-redundant part of a condition which is itself Unnecessary but Sufficient for the result." Such a condition is called an *INUS condition*. As an example, consider a house that catches fire after a short-circuit occurred in the house, with flammable material nearby. The short-circuit is said to be the cause of the fire because: (a) it is part of a true condition "short-circuit and presence of flammable material"), (b) the short-circuit is an insufficient (since flammable material is also required) but necessary (i.e., non-redundant) (since the condition is a conjunction) part of the condition, (c) the condition is unnecessary since it can be replaced by other conditions (for example, "lightning and presence of flammable material," and (d) the condition is sufficient since it, on its own, will result in the house catching fire.

David Lewis formulated a related theory of causation, based on *counterfactual dependencies* [14]. An effect is said to counterfactually depend on a cause; just in case if the cause had not occurred, the effect would not have occurred. Furthermore, A is said to be the cause of B if there is a causal chain of counterfactual dependencies linking A with B. The idea of counterfactual reasoning became both popular and influential (even if both Mackie's and Lewis's theories, of course, have been challenged by counterexamples).

A certain frustration of failing to define what "causation is" can be sensed in emotional statements by several prominent scientists. Bertrand Russell, who engaged heavily in the debate, concluded in 1913: "The law of causation, … is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." [15].[1]

By this we move on to how causation has been, and is, treated in statistics and computer science.

2.2 Statistics and computer science

We start by summarizing the relation between statistics and causation: probabilities alone cannot distinguish between cause and effect. This fact was realized by statisticians early on, but with unfortunate consequences. Karl Pearson, English mathematician and one of the founders of mathematical statistics, dismissed causation altogether in his book from 1911 [17], where he called it "a fetish amidst the inscrutable arcana of modern science." Still, as Judea Pearl remarks [18, p. 412], statistics is unable to express simple statements such as *mud does not cause rain.* There have, however, been several attempts to formulate causation in probabilistic terms. In line with the previously mentioned view of causation as a difference-maker, a cause C can be seen as something that raises the probability for an event E to occur. In its simplest form this can be expressed as

$$P(E|C) > P(E). \tag{1}$$

While this at first may look like a promising definition of a cause–effect relation, it has several problems [19]. First of all, it is symmetric in the sense that equation (1) is equivalent with

$$P(C|E) > P(C). \tag{2}$$

Hence, if C is the cause of E, then E is also the cause of C, which goes against most intuitive and formal notions of causation. Second, it does not differentiate between genuine causation and *spurious correlation*. Returning to previously mentioned ice-cream example, electricity consumption would pass as the cause of increased sales according to equation (1), with the probabilities estimated from observations. The formal reason for this incorrect inference is that a "lurking variable," the outdoor temperature, is a *common cause* or *confounding variable* to both electricity consumption and sales.[2]

Attempts to repair the definition have been made, for example, by conditioning on *background variables*:

$$P(C|E, K) > P(C|K). \tag{3}$$

For the ice-cream case, letting K be the outdoor temperature would avoid the false conclusion that electricity consumption (C) causes higher sales (E). For example, equation (3) would probably not hold for observations with $K = 25\,^{\circ}\mathrm{C}$ (or any other fixed value).[3] However, giving a general answer to which variables to include in K turns out to be far from trivial. Nancy Cartwright proposed that K should include all "causally relevant factors" [21], which may be correct but unfortunately does not provide a way to determine what these factors are. More recent approaches are suggested and discussed in ref. [22].

A simple and intuitive way to test whether electricity consumption really causes higher sales would be to try it out in reality, by some days switching on all lamps and electrical devices in the shop and observing how sales is affected. To be sure about the result, the process should of course be repeated at several random occasions. This approach has been formalized in the technique *Randomized Controlled Trials* (RCT), which was popularized by Ronald Fisher in 1925 [23]. RCT has ever since been a standard tool to find and test causal relations. The important word "randomized" stands for the *intervention*, where we some days deliberately increase electricity consumption. If this is done at random, the values of the background variables K become equally distributed between the intervened and not intervened cases [24], even if we do not know what the

[1] It is noteworthy that Russel in later work, from 1948, changed his view on causation and stated that "The power of science is its discovery of causal laws" [16, p. 308].

[2] Even if the symmetry would be enough to dismiss equation (1), the idea of causes as probability raisers is also problematic in other ways. Hesslow [20] describes an example in which contraceptive pills (C) are shown to cause thrombosis (E), besides lowering the probability of pregnancy. However, pregnancy also increases the probability of thrombosis, and the net effect may very well be that the cause C lowers the probability of its effect E.

[3] In statistical terms, this would be called "controlling for temperature" or "adjusting for temperature."

variables in K are (see Section 3.3.4 for a discussion on the role of RCTs in robotics). It is important to note that an RCT estimates causality at population-level and computes the *average causal effect*, also referred to as the *average treatment effect*. This may have very little to say about causality at the individual level. Consider, as an example, a medication that makes all men healthier, but all women sicker. An RCT with equally many men and women, would estimate the average treatment effect to zero, even if the medication causally affected every person in the experiment.

One of the two major frameworks for causal inference is the Neyman–Rubin Causal Model [25,26], where the causal effect of a binary variable C is defined as the difference between the *potential outcome* if C occurred and the potential outcome if C did not occur. However, for a specific instance, only one potential outcome can be observed, depending on whether C occurred or not. This general problem, that we cannot observe *counter factuals*, is called the *fundamental problem of causal inference* [27, p. 947]. The second major framework was popularized by Pearl with the *do-calculus* [28] (for a shorter introduction see e.g., ref. [29]). Using this formalism, we say that C causes E if [30]:

$$P(E|do(C)) > P(E). \tag{4}$$

The *do*-operator encapsulates the notion of intervention, and while $P(E|C)$ describes what value E took when a certain value of C was observed, $P(E|do(C))$ describes what values E would take if C was *set* to a certain value. The "trick" with introducing the *do*-operator is that the whole discussion about what causation "ís" gets circumvented by an axiomatic approach, in a similar way as done in for example Euclidean geometry, number theory, and probability theory.

The causal relations between a number of variables can be conveniently represented by a *Causal Bayesian Network*, which is a *Directed Acyclic Graph* (DAG), where the nodes represent variables, and the directed links (i.e., arrows) represent causal dependencies going from parent nodes to children nodes. Hence, a parent node "causes" its child node to take a specific value. Each node has an associated table with conditional probabilities given their parents. Given a DAG, the do-calculus enables inference of answers to causal questions about relations between the variables in the DAG. The DAG may be created in several ways. Conducting RCTs is one alternative but is often expensive, time-consuming, and sometimes even impossible. However, the graph structure may also be learned from observational data through *causal discovery* [7, pp. 142–154], [31], for example with the LiNGAM [32] algorithm and the PC and FCI algorithms [33], or from a formal structural causal model [29] based on prior knowledge about the problem domain. It should be noted that the directions of the causal links in the DAG cannot be learned from observational data alone without additional domain information or assumptions. For example, the LiNGAM algorithm builds on the assumption that there are no unobserved confounding variables.

To facilitate practical usage, several researchers and companies offer implementations of algorithms for statistical causal learning and inference. Microsoft has launched a software library *DoWhy*, with a programmatic interface for several causal inference methods (https://github.com/Microsoft/dowhy). The Center for Causal Discovery offers several tools, including the *Tetrad* causal discovery tool (https://www.ccd.pitt.edu/tools/). Elias Bareinboim offers *Fusion*, a tool based on Pearl's book [30] (http://bit.ly/36qUz4y). *CausalWorld* is a benchmark for causal structure and transfer learning in a simulated robotic manipulation environment [34] (https://sites.google.com/view/causal-world). Links to other software packages can be found in ref. [31], and new software is of course constantly being developed.

2.3 Cognitive psychology

Causal reasoning was, until about three decades ago, absent in cognitive psychology research. One reason may have been the general scepticism about causation in philosophy and statistics (see Section 2.1), with the result that earlier work in psychology was mostly based on correlations and associations. Reasons for this can be traced back to, at least, Hume's work in the 18th century (for an analysis of this influence, see e.g., ref. [35]).

One example of how causality enters cognitive psychology research is *causal perception* [36], which is the immediate perception of certain observed sequences of events as causal. This effect appears automatically, and is often hard to resist. For example, if the lights in the house go out exactly when you close a door, you may perceive that you caused the lights to go out, even if you know that it is totally incorrect. However, in many cases, the perception of a causal relation is correct and valuable, and causal perception can be observed already in 6-month-old infants [36, p. 4].

A historically influential model of humans causal reasoning is the *Rescorla–Wagner model* [37], which incrementally estimates the association between a conditioned and unconditioned stimuli based on observations. This and similar associative models were for a long time quite successful in modeling human behavior, but studies also indicated that human causal reasoning sometimes cannot be explained with covariation information alone [1]. A step toward causal theories were probabilistic theories such as the ΔP model [38,39], aiming at modeling how humans learn the strength between a potential cause C and an effect E. The *causal power* ΔP is given by

$$\Delta P = P(E|C) - P(E| \neg C) \tag{5}$$

where the probabilities may be estimated from observed frequency data. The ΔP model follows the previously described philosophical view of causes being difference makers (see Section 2.1). It can be seen as a variant of equation (1) and hence suffers from the same shortcomings of symmetry and common causes.

To clarify the distinction between correlation and causation: equation (5) quantifies the correlation between E and C. The two variables may very well be causally connected, but this fact cannot be established from ΔP. The reason is that both E and C may have a third variable as common cause, which may be manifested as a strong correlation, and a large ΔP. For the example given in Section 2.1, the consumption of ice-cream and electricity would have a large ΔP, even though there is no causal connection between the two variables.

The *power PC theory* [39] builds on the ΔP model and aims at modeling causal power from covariation data complemented by background knowledge. Some studies indicate that the model conforms with certain aspects of human causal reasoning. However, neither the ΔP model nor the power PC theory is seen as fully adequate to explain human causal induction [40].

Algorithms for causal inference using machine learning (see Section 2.4) have also been considered as models of human causal reasoning. Several experiments indicate that people distinguish between observations and interventions in the same fashion as Casual Bayesian Networks (see Section 2.2) [41–43]. One general concern is that these algorithms require large amounts of training data, while humans often manage to learn causal relations based on just a few observations. In cases with more than a few variables, it becomes impossible for humans to estimate all required covariations, and experiments show how humans, beside covariation data, use *mechanism knowledge* to draw causal conclusions. Studies also show that people use prior knowledge about temporal delays of different mechanisms [44]. For example, if I suddenly become nauseous, I may assume that a drug I took 2 hours ago was the cause, and not the food I ate 1 minute ago (example from ref. [1]), thereby reducing the number of covariations I need to consider. People also use cues such as spatial contiguity (the assumption that an effect is spatially close to its cause), temporal order (the assumption that a cause cannot occur before its effect) [45], and other temporal cues [46,47]. Waldmann [1,35] argues that humans employ such so called *knowledge-based causal induction* rather than Bayesian approaches based on statistical inference. Gärdenfors [48] supports this view and further argues that human causal cognition is based on an understanding of forces involved in events, and he also provides guidelines for implementations of such causal reasoning in robots.

2.3.1 Categorization of human causal cognition

Lombard and Gärdenfors [5,49] suggest a seven-grade model of the evolution of human causal cognition. The grades form a hierarchy of increasingly more complex causal skills:

1. *Individual causal understanding and tracking behavior*. Understanding the connection between a perceived motor action and the perceived resulting effect. Example: A baby learns how kicking the foot results in the foot moving in a certain way.
2. *Cued dyadic-causal understanding*. Understanding the connection between the perception of another human's actions and the effect they have.
3. *Conspecific mind reading*. Understanding how another human's desires, intentions, and beliefs lead to different kinds of actions. Example: If a person fetches a glass of water, I infer that the person is thirsty.
4. *Detached dyadic-causal understanding*. Understanding the connection between another human's actions (non-perceived) and the effect (perceived). Example: You perceive the tracks of a person in the snow, and conclude that somebody's presence in the past is the cause of the tracks.
5. *Causal understanding and mind reading of non-conspecifics*. Example: If I see an animal, or even the tracks of an animal, I may be able to infer the mental state of the animal.
6. *Inanimate Causal Understanding*. Understanding causal relations between inanimate objects. Example (from ref. [49]): by observing an apple falling from a tree at the same time as a gust of wind, I infer that the wind caused the apple to fall.
7. *Causal Network Understanding*. Understanding how nodes in one causal network is linked to nodes in a causal network in another domain. Example (from ref. [49]): once I learned that wind can cause an apple to fall, I may understand that wind can also cause other things to fall or move.

While the seven grades are presented as a model of human evolution, we find them valuable also to describe an "evolution" of robots with casual cognition and will build on them for that purpose in Section 3.1.

2.4 Causation in machine learning

In machine learning, the recent progress in deep learning has been achieved without any explicit notion of causation. The results from applications in speech recognition [50], natural language translation, and image processing [51] can be seen as the victory of data over models, as claimed already in 2009 by Halevy, Norvig, and Pereira in their article *The Unreasonable Effectiveness of Data* [52]. A common view seems to have been, and to a large extent still is, that most problems can be solved by providing sufficient amounts of data to sufficiently large computers running sufficiently large neural networks. As an extreme example, the current state-of-the-art model for language processing, GPT-3 [53], was trained with almost 500 billion tokens and contains 175 billion parameters. Given that the size of the model is of the same order of magnitude as the modeled data, the comparison with a giant lookup table is not far-fetched. Nevertheless, when GPT-3 was introduced, it outperformed previous state-of-the-art on a large range of natural language processing tasks, including translation, question-answering, and generation of news articles.

Despite the unquestionable success of such purely data-driven approaches, critical voices have expressed concerns about their true "ĭntelligence." Judea Pearl, one of the pioneers in AI and causal reasoning, argues that deep learning is stuck on the level of associations, essentially doing advanced curve fitting [8,54] and clustering, and only modeling correlations in data. As such it works for predictions, answering questions such as "What will the outcome of the election be?" However, it cannot answer causal questions, like "Would we have won with a different financial policy?," or "How can we make more people vote for us?" The former question requires *counterfactual* reasoning about possible alternatives to events that have already occurred. An answer to the latter question provides an action for *control*, such as "By lowering the interest rates." Both counterfactuality and control are out of reach for methods using observational data only.

The reliance on predictive algorithms in machine learning, combined with a dramatic increase in the use of machine learning for real-world tasks, is for these reasons seen a serious problem [55].

Léon Bottou, a leading researcher in deep learning, points to the fundamental problem of machine learning "recklessly absorbing all the correlations found in training data" [56]. This leads to sensitivity to spurious correlations stemming from data bias,[4] and a failure in identifying causal relations, for example, between features and classification of objects in images. Examples of such failures, and an argumentation for a general need to incorporate causation in models of human-like intelligence can, for example, be found in ref. [4]. The specific connection between causality and generalization is analyzed in ref. [58].

Important steps have also been taken to incorporate both causal inference and causal reasoning in machine learning, in general (see refs. [7,59] for overviews), and deep learning, in particular [4,60–63]. Two specific techniques for supervised learning will be described in more detail. The Invariant Risk Minimization (IRM) paradigm [56] addresses the observed problems with "out-of-distribution" generalization in machine learning. To train neural networks that also function with previously unseen data, the training data are first divided into *environments*. The loss function is extended by a term penalizing performance differences between environments. This promotes networks that are invariant over environments, in a way that is fundamentally linked to causation. The Invariant Causal Prediction (ICP) approach [64] is a technique to identify features that are causally connected to a target variable (as opposed to only being connected through correlation). Combinations of features are used to build separate models for data from several different environments. Combinations that work well over all environments are likely to contain features that cause the target variable. The intersection of several such successful combinations are regarded as the overall causally relevant features.

3 Causation in robotics

In this section, we present an analysis of the role of causal reasoning in robotics, organized in two parts. The first part is a novel categorization of robot causal cognition, inspired by the categorization of human causal cognition in ref. [5], previously described in Section 2.3.1. The latter describes a hierarchy of seven *grades* of causal skills, with humans mastering all grades, and animals

[4] In this case, bias refers to *sampling bias* [57], causing a skewed distribution for some feature in the training data.

only certain grades, all according to their stage in evolution. Our categorization defines eight *categories* that relate to these grades; however, with several important differences as discussed in Section 3.2. The categories describe causal cognition mainly at the sense–plan–act level.[5] The second part of the analysis describes the usage of causation in robotics, beyond the sense–plan–act level covered in the first part.

Throughout the section, we review related earlier work in robotics and also identify connections between the previously described work in philosophy, statistics, computer science, and cognitive psychology.

3.1 Causation for sense–plan–act

Our categorization comprises eight categories of robot causal cognition at the sense–plan–act level. The underlying causal mechanisms are illustrated in Figure 1, with a robot, a human, and the (rest of the) world, affecting themselves and each other causally, illustrated by the arrows, and with numbers referring to the corresponding categories. The categories are illustrated in more detail in Figure 2. An interacting human's internal causal processes are included since they are important for the robot's causal reasoning. The top row shows the human *Mind* initiating (i.e., causing) *Motor Forces* (e.g., muscle activations) that result in *Motor Actions*, (e.g., limb motions). Motor Actions may in turn cause *World Effects* since the human is situated in the environment. The robot's mind is modeled in a similar way, with additional World Effects added to model non-animate causal relations. For both the human and the robot, the term "mind" denotes the mechanism that, for one reason or another, causes Motor Forces. In a deterministic world, the activation of such a mechanism, of course, also has a cause. The usage of the word "mind," with its associations to free will and non-determinism, indicates that we here regard the mind as "the ultimate cause."

In the figure, all causal relations are illustrated by thick arrows. Inference of causes is illustrated by thin arrows pointing toward the inferred entity.

Categories 1–3 have to do with *Causal Learning*, i.e., how a robot learns causal relations involving itself, interacting humans, and the world. Categories 4–6 have to do with the important cases of inference of causes related to an interacting human. Categories 7–8 have to do with how the robot decides how to act, with more or less sophisticated usage of causal reasoning.

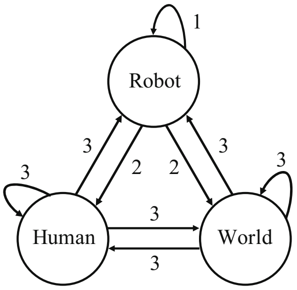

Figure 1: Causal mechanisms between a robot, a human, and the (rest of the) world. The arrows go from cause to effect, and the numbers refer to the corresponding category further described in the text.

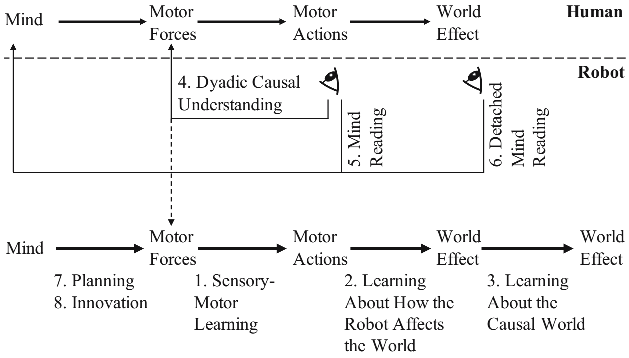

Figure 2: The functions of the eight categories of robot causal cognition. Categories 1–3 refer to learning of causal relations (bold arrows). Categories 4–6 refer to inference of causes (thin solid arrows) related to an interacting human, while categories 7–8 refer to how the robot decides how to act.

Category 1. Sensory–motor learning

This causal skill is about learning a mapping from Motor Forces to Motor Actions (Figure 2). The former are activated by the robot's Mind, as a result of some decision activity, to be further described as categories 7 and 8. This causal mechanism is also illustrated by arrow 1 in Figure 1.

Sensory–motor learning is strongly related to *motor babbling*, which is the method supposedly used by infants to learn the mapping between muscle movements and the resulting change in body pose and limb positions [65]. Motor babbling has also been suggested as a plausible model of humans' acquisition of more complex behaviors, such as movements along trajectories for avoiding

[5] In this context the term planning refers to all sort of low- or high-level mappings between sensing and acting and not only planning in its original meaning in robotics.

obstacles or movements directed to grasp objects [66]. Motor babbling has been fairly well explored in robotics [67–69] with the same aims. The typical setup is that the robot learns a mapping between randomly explored joint settings and the resulting perceptual inputs, from touch sensors or visual sensors. For complex robots with lots of joints, the "curse of dimensionality" calls for special solutions [70]. This mapping is strongly related to the "the kinematic equations," which map joint settings to the configuration of, for example, a robot arm. In ref. [71], a robot learns the causal relations between moving the arm, touching an object, and the object visually moving. The results show how causal dependencies improve scene understanding by discriminating real objects from visual artefacts.

In causal terms, motor babbling is related to reinforcement learning (see Section 2.2), where the robot partly executes random actions and receives reward depending on whether the actions led to (i.e., caused) an advantageous state or not.

Category 2. Learning about how the robot affects the world

This category has to do with how a robot learns how its Motor Actions have a causal World Effect. See Figure 2 and also the arrows labeled 2 in Figure 1. One example is the mechanism by which a computer is switched on by pressing a certain button, and another is how a service robot may make its user happy by suggestion her to read a book. The previously described RCT technique (see Sections 2.2 and 3.3.4) provides a general approach for such learning, and a similar experimental approach was applied in ref. [72], where a Baxter robot learned tool affordances (what tools can be used for). The robot first observed demonstrations of a block being pushed with a hoe tool by a human teacher. The previously mentioned PC algorithm (Section 2.2) for causal learning was then used to create an initial structural causal model in the form of a DAG [18]. In order to orient the edges of the DAG, the robot then intervened by forcing one variable to take a new value and observing the result. The difference between the two resulting interventional distributions was then estimated by a nonparametric test. With additional interventions, the robot added robot actions as nodes to the DAG.

Category 3. Learning about the causal world

This category has to do with how a robot learns how the world evolves through causal mechanisms. One example is learning physical laws like gravitation and inertia. This category also includes how inanimate objects and humans affect themselves, each other, and the robot. One such example is a robot companion learning how ice on the road sometimes makes the human slide. Another example is a robot learning that the user tends to forget charging it after having watched TV late at night.

In Figure 2, this category is illustrated by the arrow between the two World Effect items, and in Figure 1 by the arrows labeled 3. Humans often find such causal relations by observing co-occurrences of events and computing measures such as the causal power ΔP in equation (5). Robots may of course do the same, with the same associated risk of taking correlation for causation.

An experimental approach addresses this risk and was explored in refs. [73,74], where the authors describe an iCub humanoid robot that learns simple causal relations related to Archimedes' principle. For example, the robot learns that the amount of water an object displaces is independent of the color of the object. This insight is reached by cumulative learning of causal relations through experimentation. Based on its experiences, the robot chooses actions it believes will cause the desired response, and step-by-step updates its causal hypotheses.

Just like categories 2 and 3 is closely related to the techniques for statistical causal inference described in Section 2.2. Any real-world application has to deal with complex causal relations that go beyond single causes having single effects. In the real world, causes are often neither sufficient nor necessary, and the intertwined relations are best expressed with techniques such as causal Bayesian networks. As described earlier, several algorithms and software tools have been developed, both to create such networks and to conduct causal inference with them.

Category 4. Dyadic causal understanding

This category has to do with how observations of another agent's motor actions are used to infer the cause of the motor action. Humans accomplish this task with the aid of mirror neurons, which are a specific kind of neurons that respond to both observed and self-produced actions [75]. By observing another human's motor action, for example kicking a ball, an observing human may perceive unintentional movements of the leg [76]. Mirror neurons have also been suggested to play a role in human empathy [77]. For example, if the kicker grimaces after hitting the ball, the observing human may also the "feel" pain. For humans, some of these mechanisms are encoded in innate mirror neurons mechanisms such that they do not have to be learned [78].

Category 4 refers to this type of causal inference mechanism for a robot observing an interacting human. The mechanism is illustrated in Figure 2, with the arrow going from the observing robot eye to the human's Motor Forces, which are inferred as the cause of the perception.[6] The dotted arrow illustrates the mirroring to the robot's Motor Forces.

Earlier related work include refs. [79,80], in which mirror neurons are implemented to support action interpretation in imitation learning and also for goal-dependent action-understanding [81].

Category 5. Mind reading

There is large body of robotics research addressing how a robot reads parts of an interacting human's mind. The terms "mind reading," "theory of mind," "mentalizing," and "perspective-taking" are commonly used to denote this process [82,83]. A common focus is the human's ongoing and planned actions and goals and also the underlying reasons for the human's actions. In *plan recognition*, *intention recognition*, and *behavior recognition*, a human's actions are used to infer a plan, intention, or behavior that is assumed to govern the human's behavior. A robot may utilize this knowledge in many ways, for example, to support the human or to improve collaboration. Example: A robot at an elder care facility observes a human taking out a bottle of juice from the refrigerator and infers that the human is thirsty and intends to drink. The robot therefore recommends the human to first test the blood sugar level. Note that this inference is susceptible to confounding variables (see Section 2.2). One do not have to be thirsty to take a bottle of juice from the refrigerator. Other possible causes may be to offer someone else juice, to throw away old juice, or to clean the refrigerator. Hence, correlation-based techniques are not sufficient, and context may be necessary to make the correct inference [84].

This type of inference is addressed in category 5 and is illustrated in Figure 2, with the robot's eye observing the human's Motor Actions. Based on these observations, the human's plan, intention, or behavior is inferred as a possible cause.

The inference of a human's mind is typically based on an assumption of rationality on behalf of the observed human [85], such that the human's actions are seen as being *caused* by the plan, intention, or behavior. Only rarely is this relation to causation mentioned in published research. One exception is ref. [86], in which intent recognition and plan adaption is integrated for collaborating human–robot teams. Causal link analysis is used to narrow down the possible options that the human has, thereby improving intent recognition as well as adaption of the robot's actions.

Mind reading is relevant for several techniques where human operators teach robots skills. The most common such techniques are *learning from demonstration* (LfD) [87,88] and *imitation learning* (IL) [89]. In LfD, a human operator typically remote controls the robot, and the robot tries to infer the intention of the operator. In IL, the human performs the task while the robot observes the human and then tries to infer her intention. In both cases, the aim is to make it possible for the robot to repeat the demonstrated behavior, also in new situations with different conditions than during the learning phase. The human is usually assumed to act rationally, in the sense that each action is performed because it *causes* a specific wanted effect. This means that both LfD and IL are causal by definition.[7] However, this does not mean that causal inference techniques are used, or that causation is even mentioned in published research on LfD and IL. Instead, most published work rely on observed correlations between sequences of actions and goals, intentions, or behaviors. Some important exceptions are briefly reviewed below.

In refs. [90,91], LfD is used to teach a robot arm to follow trajectories demonstrated by a human teacher. In order to generalize, causal reasoning is used to identify which of three pre-defined user types (differing in degrees of cautiousness in motion) a demonstrated trajectory belongs to. The authors of ref. [92] address a similar problem in a multimodal setting. Using RGB images, sound spectra, and joint angles as inputs to a deep neural network, a robot may learn cross-model causal dependencies. Despite this claim made in the article, it is unclear how the correlation and causation are distinguished from each other.

[6] Strictly speaking, the result of the inference does not map to the human's Motor Forces but rather to the robot's model of the human. For simplicity, this distinction is not made in the figure. The same simplification is made in the illustration of categories 5 and 6.

[7] Note that the robot does not aim at learning causal relations during LfD or IL but rather assumes that the human acts according to causal principles, while the robot tries to learn the human's goal. For that reason, we view LfD and IL as examples of a robot conducting causal inference rather than causal learning.

In ref. [93,94], causal reasoning is explicitly utilized for IL. The *Parsimonious Covering Theory* (PCT), a formal computational model of cause-effect reasoning, is extended and used to infer sequences of human intentions that explain demonstrated actions in a causal fashion. This should be compared to imitating the observed actions literally, which is often the approach taken in IL. Another approach that combines causal modeling of the environment with IL is presented in ref. [95].

In ref. [96], the authors described an approach to understand the causal structure of a demonstrated task in order to find which variables cause what other variables to change. A robot simulator was used to generate data for a demonstrated pick and place task, and Tetrad, a well-known computer program [97] for causal analysis, was used to identify irrelevant variables in the demonstrated task.

The authors of ref. [98] describe, as an example, how IL can be used to teach an autonomous car when to brake, based on images shot by a camera mounted inside the car. The idea is that the robot should learn to brake when a pedestrian appears in certain locations in the image, and data for this learning are collected with a human driver operating the brakes. However, this can go terribly wrong if the braking indicator lamp of the car is visible in the images. The learning algorithm may in such case learn to activate the brakes when the lamp is lit, which is a clear case of how cause and effect may be mixed up by correlation-based approaches. The authors denote the mistake as *causal misidentification* and describe it as a naturally occurring and fundamental problem in IL and generally in machine learning systems deployed in the real world. Two interventional strategies are proposed to overcome causal misidentification one based on environmental rewards and the other on human experts providing additional input.

In ref. [99], the authors present a robot that learns a cloth-folding task after watching a few human demonstrations. The robot reproduces the learned skill and manages to generalize the task to other articles of clothing. While the used hard-coded graph represents causal relations, the inference mechanism only relies on correlations observed during the demonstrations.

In ref. [100], the design of an indoor cleaning robot is described. In order to automatically switch between four operation modes (sweeping, absorption, grasping, and erasing), the mapping from perceived garbage attributes (e.g., size, solid/liquid, flat/non-flat) and the appropriate operation mode is built by modeling observed manual decisions with causal inference techniques. A causal Bayesian network in the form of a DAG is constructed using the *DoWhy* toolbox (see Section 2.2).

Category 6. Detached mind reading

Sometimes the robot only perceives the effect of a human's action, which in itself is not perceived. Category 6 refers to the causal skill of inferring the action in such situations. Example: the robot sees an empty juice bottle on the kitchen counter and concludes that the human has been drinking juice.

This type of inference requires of the robot to handle two mental representations at the same time: the current perceptual state and the mind of the human in the past [49]. The inference is illustrated in Figure 2 in the same way as for category 5, but with the robot's eye now observing the World Effect. Based on this information, the human's plan, intention, or behavior is inferred as a possible cause.

The inference of the cause of a perceived effect answers the question *why* the effect was observed. This may also be seen as generation of an explanation for the effect. The authors of ref. [101] describe methods for inference of actions not performed by the robot (the so-called exogenous actions) from observations. The specific techniques used include KRASP, a system for Knowledge Representation for Robots using ASP, and are based on Answer Set Prolog (ASP) that allows representations of causal relations. Generation of explanations is also described as a part of understandable robots in Section 3.3.1.

Category 7. Planning

By *Planning* we here refer to all processes by which a robot generates single, or sequences of, actions or behaviors that lead to the robot's currently set goal. In AI and robotics, this goes under names such as *Action Selection*, *Arbitration*, and *Planning*. Planning is inherently causal since the chosen actions or behaviors are expected to cause known effects and not merely be correlated to them. This was noted already in early research in planning [102] and is still sometimes mentioned explicitly, even if it most often is an implicit assumption.

The mappings of actions to effects are either hard-coded or learned, based on non-causal [103] or causal [104] approaches. The previously described categories 1–3 are all relevant for such causal learning. For example, the robot may hit a door by driving against it (category 1), a door may swing open if being hit (category 2), and the opening of the door may cause people in the room to turn toward the door (category 3).

In ref. [105], causal reasoning is applied for planning and coordinating multiple cleaning robots. Causal laws are formulated in the action description language *C*+ [106], and a causal reasoner *CCALC* [107] is used for planning. A similar approach is taken for task planning in ref. [108], where the action domain and the planning problem are formulated as *causal theories* [107] such that a *causal model* corresponds to a planning solution which is found using sophisticated solvers.

Category 8. Causal innovation

We do not define this category very precisely, but see a need to go beyond category 7, and cover what is often denoted "common sense," or "human level intelligence." Such abilities involve a range of cognitive skills beyond causal reasoning, such as identification of similarities between problem domains, generalization of observations, imagination, fantasy, and curiosity. However, many of these skills relate to answering counterfactual questions (see Section 2.2) like "what would have happened if I had knocked on the closed door instead of driving against it?," and "what would have been the advantages that?." This kind of causal reasoning is placed at the top of the three-level causal hierarchy by Judea Pearl, above the interventional and associational levels [8]. Counterfactual reasoning is sometimes mentioned as a necessity for truly intelligent robots but has rarely been applied in published research. In ref. [109], counterfactual reasoning is used to find ways to modify a scenario such that a robot controller manages a given task. New "hallucinated" (counterfactual) data are generated by modifying the observed scenario data such that a simulated robot manages the task.

3.2 Discussion on the categorization

Below we summarize the eight identified causal cognition categories. For ease of comparison, the most similar grade in the categorization of human causal cognition in Section 2.3.1 is given within parentheses. The related technique listed for each category refers to a causal concept or technique from robotics or one of the other reviewed scientific fields. As such they should only be seen as examples.

1. *Sensory Motor Learning* (1) – Related technique: Motor Babbling
2. *Learning How the Robot Affects the World* (-) – Related technique: Interventions
3. *Learning About the Causal World* (6) – Related technique: Experiments
4. *Dyadic Causal Understanding* (2) – Related technique: Mirror Neurons
5. *Mind Reading* (3,5) – Related technique: Intent Recognition
6. *Detached Mind Reading* (4) – Related technique: Statistical Causal Inference
7. *Planning* (-) – Related technique: Planning Algorithms
8. *Causal Innovation* (7) – Related technique: Counterfactual Reasoning

It is interesting to note that all seven grades of human causal cognition can be associated with existing techniques within robotics. In the other direction, categories 2 and 7 have no direct correspondence in the grades. However, it can be argued that category 2 is included in grade 1 and category 7 in grade 7.

We found earlier work in robotics that addressed all eight categories. However, for all but category 5, we found very few papers, in particular papers where the connections to causal concepts and reasoning are explicitly mentioned.

3.2.1 Comparison with the categorization of human causal cognition

There are several differences between our categorization of robot causal cognition and the categorization of human causal cognition described in Section 2.3.1. One difference concerns the distinction between human and non-human animals. For human cognition, this distinction is important since inference related to humans and non-humans is believed to have taken place separately during evolution. Humans and non-humans were therefore placed in separate grades. Furthermore, inference of humans was placed in lower grades, reflecting that these skills are believed to have evolved first. We choose to merge inference of humans' and non-human animals' minds into the same category 5 and also to consider other robots as possible agents. The main reason is that the conditions for causal inference is very similar. For example, an assumption of rationality should apply to the inference of both robots' and humans' intentions.[8]

Regarding causal understanding of agents versus non-agents, we keep the distinction made in ref. [5]. Indeed, humans do not have any mirror neurons that model, for example, stones. And stones are usually not

[8] However, it would obviously be much more efficient if robots directly communicate goals and planned actions to each other, thereby eliminating the need for mind reading altogether. This is of course only possible if the robots "speak the same language," and are programmed to share information with each other.

regarded to be equipped with minds that are possible to read. Causal understanding of non-agents therefore requires other techniques, such as causal models and experiments, as described in category 3.

Another difference is that we characterize categories as related to either causal learning or causal inference. This is a natural choice since learning and inference most often are realized with very different techniques in a robot. Furthermore, this separation is also common in cognitive psychology (even if there, as always, are differing opinions [110]).

We also introduce two new categories that we believe are crucial for causal robots. Category 2 has to do with the robot learning how its actions affect the world, and category 7 has to do with planning, in the broad sense, covering all sorts of inference on how to act.

As a consequence of our ordering of categories, the hierarchical ordering of the seven grades in ref. [5] is not maintained. We do not see this as a problem, since the development of causal robots should not be expected to follow a linear trajectory (the same is actually also assumed for human's evolutionary development of causal skills [111]).

3.3 Causation beyond sense–plan–act

Robotics is more than designing functionality for robots to sense, plan, and act, and causation also plays an important role in these other areas and aspects of robotics. A few examples are given below, even if the list undoubtedly could have been made much longer.

3.3.1 Understandable robots

The area *understandable robots* [112], sometimes denoted *explainable robots*, or *explicable robots*, receives increased attention and has several connections to causation. An understandable robot should, among other things, explain the reasons for its actions to interacting humans. Such explanations are causal in most cases [113], or maybe even in all cases [114], and have been shown to affect humans' trust in robots [115]. For other examples of research on generation of explanations in robotics, see refs. [101,116–119].

An understandable robot should communicate not only the reasons for its actions but also the reasons why it is *not* acting as expected, and also its goals, beliefs, desires, limitations, and capabilities [112]. All communication for understandability should be governed by a causal analysis of the consequences of informing or not informing an interacting human. For example, a robot assisting a pedestrian to cross a street should inform about an approaching car if there is a risk of collision, and the human most likely is unaware of the car. Furthermore, the robot has to decide when, to whom, and on which level of details it should communicate [120]. All such decisions require a causal analysis to assess available alternatives.

3.3.2 Robot ethics

As robots are getting increasingly ubiquitous and autonomous, several new ethical questions arise. Many of them are intimately connected to causation.

One major question is how to design robots that behave ethically according to the norms of the society in which they are deployed. One approach is to implement rules or principles that govern the robot's behavior in the desired way. This requires of the robot to predict the consequences of possible actions, and to generate plans that leads to desired goals. For this, the robot needs causal skills from, at least, categories 2 and 7 (and in the general case also from all the other categories).

A specific problem is how to ensure that robots behave in an unbiased and non-discriminatory way. For example, a waiter robot should treat male and female customers equally, in the sense that sex in itself should not affect the order in which the robot serves customers. This is a general problem when big data is being used to train decision support systems, since they often inherit bias in the data [57]. Approaches to deal with this have only recently been addressed from a causal perspective [121].

Another major question concerns whether a robot can be regarded as responsible for its actions. A robot's, or more generally, an agent's moral responsibility is often described with two components: causal responsibility and intention (see e.g., ref. [122]). Being causally responsible for an event here means to be the cause of the event, often in the sense of making a difference (see Section 2.2). In criminal law, this is often posed as a counterfactual question: "but for the defendant's action, would the victim have been harmed as she was?" [123]. Intention, the second component of moral responsibility, is also connected to causation. For a robot to have the intention to cause an event, it should, at a minimum, be aware of how its actions and non-actions may cause the event. Ultimately, robots that we consider as morally responsible (as such a debatable question [124–126]) will probably need causal competence from all of our eight identified categories. Hence, the robots must step up to the highest rung of Pearl's "causal ladder" [30] in order to master the necessary counterfactual reasoning.

3.3.3 Machine learning for robotics

Machine learning has always been applied in robotics, for example, to develop object identification and navigation functionality. As such, the problems with current non-causal machine learning mentioned in Section 2.4 also are highly relevant for robotics.

Machine learning also plays a major role as a way for robots to learn how to act. We have previously mentioned LfD and IL as two common learning paradigms. Another paradigm is *reinforcement learning* (RL), which has clear connections to both robotics and causation. In its traditional form, RL is a technique by which a robot may form optimal behaviors through trial and error with the environment (see, for example, ref. [127]). The robot learns what to do in each state by getting positive or negative reinforcement after each action. The resemblance with interventions (see Section 2.2) is clear: the robot learns causal links between actions and reward, and not merely correlations between observational data. Since in reality a complete state may not be observable, one can argue that RL is not entirely model-free [128], and that some level of causal knowledge is necessary, and often implicitly introduced, to estimate hidden states from partial observations, or to design simulators used to generate training data for RL. In *causal reinforcement learning*, causal knowledge is explicitly introduced, for example, as *causal diagrams*, with the effect that the dimensionality of the learning problem is reduced [129].

3.3.4 Robotics research methodology

Causation is important, not only to equip robots with causal skills, but also as part of common robotics research methodology. Randomized Controlled Trials (RCTs) have already been mentioned several times in this article, as a technique that a robot can use to deal with confounding variables in causal inference. However, RCTs have a very prominent role also as a way to draw causal conclusions regarding the effect of certain design choices or to assess how external factors influence a robot's function and interaction with humans. *User studies* is probably the most common approach to verify hypotheses on how humans perceive interaction with robots, and is often organized as an RCT (see e.g., ref. [130] for a thorough description). An *independent variable* refers to a "causal event that is under investigation" [131], for example, the gender of a robot's voice. The independent variable is hypothesized to affect the *dependent variable* (also denoted *effect*), for example, user satisfaction. A number of *test subjects* are then randomly assigned to groups, where each group "receives different treatments," meaning that the independent variable takes different values for each group. In the example, the test participants in one group would interact with a robot with a male voice, while the others interacts with a robot with a female voice. Due to the random assignment to groups, the effect of confounding variables is reduced such that one can draw causal conclusions from observed correlations between the independent and dependent variable. In the example, a conclusion from the experiment could be that a robot with a male voice makes people less patient.

Path analysis, a special case of *structural equation modeling* [132], is a more sophisticated method for causal analysis of experiments. In ref. [133], structural equation modeling was used to study the relation between touch sensations (if the robot is "soft") and personality impressions (if the robot is "talkative," etc.).

The general importance of causal learning for robots is receiving increased attention. In ref. [34], a simulator for development and testing of algorithms for causal learning with robots is presented. The simulator implements basic physics and allows a user to define block-world manipulation tasks with a simulated robot and to conduct interventions in the causal structures.

4 Summary and conclusions

We presented a two-part analysis of the role of causation in robotics. The first was a novel categorization of robot causal cognition at the sense–plan–act level, summarized in Section 3.2. We also identified connections between each category and causal concepts and techniques in philosophy, psychology, and computer science/statistics. For most categories, very few related publications were found. The exception was category 5: Mind reading, where some published work on robot learning recognizes the importance of causal reasoning. However, causation is rarely mentioned explicitly.

We also discussed how robotics research beyond the sense–plan–act level depend on causal concepts and assumptions, e.g., understandability, machine ethics, and research methodology. Also in these cases, causation is rarely explicitly mentioned in the research articles. One reason for this, mentioned in ref. [134], may be that most theoretical development in causation has been based on *structural equation modeling*, originating in econometrics [132]. This has led to a focus on *population-level* models [135], describing, for example, general causal relations between unemployment benefits and unemployment levels in a society [136]. While such relations may be relevant in econometrics, they do not address the causal relation between a specific unemployed individual receiving specific unemployment benefits. This kind of relations is better handled with *individual-level* models such as the *sequences of mechanisms* [134], and other approaches [117,137], of which unfortunately there are not very many. For a robot, individual-level inference is probably more relevant than population-level, at least until robots become considerably more advanced than they are today. Another reason may be that the causal relations of relevance in robotics often involve high-level concepts, while available data typically comes from low-level sensors like laser scanners and ultra-sonic sonars. Higher-level perception, such as object classification, and recognition of faces, intentions, and emotions, would be more appropriate in this respect, but research in these areas still struggle with generality, and robustness in the real-world scenarios robots typically operate in. This means that high-level perception cannot be easily applied in complex cognitive computations such as causal reasoning.

The fact that causation is so rarely explicitly mentioned in published robotics research, has several possible consequences. The most serious would be if causality is simply not taken into account, leading to mistakes like interpreting correlation in data as causation. This risk is particularly high when data-driven machine learning is involved. As sub-systems building on machine learning are integrated in robots, this risk will also be an issue for robotics. Maybe less serious, but still highly undesirable, would be if causation is recognized as an important part of a problem, but connections to other research areas, such as philosophy, cognitive psychology, statistics, and computer science, are not identified and fully exploited. This emphasizes the importance of a cross-disciplinary approach to fully integrate causation into robotics research.

It is our hope that this article fills a gap by clarifying the broad and deep connections between causal reasoning and robotics and by pointing at the close connections also to other research areas. At best, this will contribute to a "causal revolution" [54] also in robotics.

Acknowledgements: The author wishes to thank Xavier de Luna and the two anonymous reviewers for extremely valuable feedback and suggestions for improvements to the article.

Funding information: This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721619 for the SOCRATES project.

Conflict of interest: The author states no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no data sets were generated or analysed during the current study.

References

[1] M. R. Waldmann, "Causal reasoning: an introduction," in *The Oxford Handbook of Causal Reasoning*, M. R. Waldmann, Ed., Oxford, UK: Oxford University Press, 2017, pp. 1–9.

[2] M. Stuart-Fox, "The origins of causal cognition in early hominins," *Biology & Philosophy*, vol. 30, pp. 247–266, 2015.

[3] J. Bering and D. Povinelli, "The mentality of apes revisited," *Curr. Dir. Psychol. Sci.*, vol. 11, no. 4, pp. 115–119, 2002.

[4] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav. Brain. Sci.*, vol. 40, p. e253, 2017.

[5] M. Lombard and P. Gärdenfors, "Tracking the evolution of causal cognition in humans," *J. Anthropol. Sci.*, vol. 95, pp. 219–234, 2017.

[6] P. Gärdenfors and M. Lombard, "Technology led to more abstract causal reasoning," *Biology & Philosophy*, vol. 35, p. 23, 2020.

[7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, Cambridge, MA, US: MIT Press, 2017.

[8] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Commun. ACM*, vol. 62, pp. 54–60, 2019.

[9] P. Spirtes, "Introduction to causal inference," *J. Mach. Learn. Res.*, vol. 11, pp. 1643–1662, 2010.

[10] J. Woodward, "A philosopher looks at tool use and causal understanding," in *Tool Use and Causal Cognition*, T. McCormack, C. Hoerl, and S. Butterfill, Eds., Oxford: Oxford University Press, 2011, pp. 18–50.

[11] R. L. Anjum and S. Mumford, *Causation in Science and the Methods of Scientific Discovery*, Oxford, UK: Oxford University Press, 2018.

[12] J. L. Mackie, "Causes and conditions," *American Philosophical Quarterly*, vol. 2, no. 4, pp. 245–264, 1965.

[13] J. L. Mackie, *The Cement of the Universe: A Study of Causation*, Oxford, UK: Oxford University Press, 1974.

[14] D. Lewis, "Causation," *J. Philos.*, vol. 70, no. 17, pp. 556–567, 1973.

[15] B. Russell, "I.-On the notion of cause," *Proceedings of the Aristotelian Society*, vol. 13, pp. 1–26, 1913.

[16] B. Russell, *Human Knowledge: Its Scope and Limits*, New York, USA: Simon and Schuster, 1948.

[17] K. Pearson, *The Grammar of Science*, Adam and Charles Black, 3rd ed., 1911.

[18] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, Cambridge, 2009.

[19] M. Benzi, "Probabilistic causation," in *Proceedings of the XVI Summer School in Philosophy of Physics, Forecasting the Future: Epistemology and Empirical Sciences*, 2014.

[20] G. Hesslow, "Two notes on the probabilistic approach to causality," *Philos. Sci.*, vol. 43, 1976.

[21] N. Cartwright and E. McMullin, "How the laws of physics lie," *American J. Phys.*, vol. 52, pp. 474–476, 1984.

[22] J. Häggström, "Data-driven confounder selection via markov and bayesian networks," *Biometrics*, vol. 74, no. 2, pp. 389–398, 2018.

[23] R. Fisher, *Statistical Methods for Research Workers*, Edinburgh, UK: Oliver and Boyd, 1925.

[24] N. Cartwright, "What are randomised controlled trials good for?," *Philos. Stud.*, vol. 1, pp. 59–70, 2010.

[25] J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: essai des principes," *Roczniki Nauk Rolniczych X*, vol. 5, pp. 1–51, 1923.

[26] D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Edu. Psychol.*, vol. 66, pp. 688–701, 1974.

[27] P. W. Holland, "Statistics and causal inference," *J. Am. Stat. Assoc.*, vol. 81, no. 396, pp. 945–960, 1986.

[28] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[29] F. Dablander, "An introduction to causal inference," Feb 2020, Psy ArXiv: DOI: https://doi.org/10.31234/osf.io/b3fkw.

[30] J. Pearl and D. Mackenzie, *The Book of Why - The New Science of Cause and Effect*, Allen Lane, New York, US: Basic Books, 2018.

[31] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Front. Genet.*, vol. 10, p. 524, 2019.

[32] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, pp. 2003–2030, 2006.

[33] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, vol. 81, New York, USA: Springer, 1993.

[34] O. Ahmed, F. Tråuble, A. Goyal, A. Neitz, M. Wüthrich, Y. Bengio, et al., "Causal World: A robotic manipulation benchmark for causal structure and transfer learning," arXiv:2010.04296v1 [cs.RO], 2020.

[35] M. R. Waldmann, "Knowledge-based causal induction," in *The Psychology of Learning and Motivation*, D. R. Shanks, K. J. Holyoak, and D. L. Medin, Eds., San Diego, US: Academic Press, 1996, pp. 47–88.

[36] D. Danks, "The psychology of causal perception and reasoning," in *The Oxford Handbook of Causation*, H. Beebee, C. Hitchcock, and P. Menzies, Eds., Oxford, England: Oxford University Press, 2009.

[37] R. A. Rescorla and A. R. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Theory and Research*, A. H. Black and W. F. Prokasy, Eds., New York, US: Appleton-Century-Crofts, 1972, pp. 64–99.

[38] H. M. Jenkins and W. C. Ward, "Judgment of contingency between responses and outcomes," *Psychol. Monogr.*, vol. 79, no. 1, pp. 1–17, 1965.

[39] P. W. Cheng, "From covariation to causation: A causal power theory," *Psychol. Rev.*, vol. 104, 1997.

[40] J. Tenenbaum and T. Griffiths, "Structure learning in human causal induction," in *Proceedings of Advances in Neural Information Processing Systems 13* (*NIPS 2000*), 2000.

[41] H. Lu, A. Yuille, M. Liljeholm, P. Cheng, and K. Holyoak, "Bayesian generic priors for causal learning," *Psychol. Rev.*, vol. 115, pp. 955–84, 2008.

[42] A. Coenen, R. Rehder, and T. Gureckis, "Modeling active learning decisions during causal learning," in *Proceedings of the 1st Multidisciplinary Conference on Reinforcement Learning and Decision Making*, Princeton, NJ, 2013.

[43] Y. Hagmayer, S. Sloman, D. Lagnado, and M. R. Waldmann, "Causal reasoning through intervention," in A. Gopnik and L. Schulz, Eds., *Causal Learning: Psychology, Philosophy, and Computation*, Oxford, UK: Oxford University Press, 2007, pp. 86–100.

[44] M. Buehner and J. May, "Knowledge mediates the timeframe of covariation assessment in human causal induction," *Thinking & Reasoning*, vol. 8, pp. 269–295, 2002.

[45] P. White, "Singular clues to causality and their use in human causal judgment," *Cogn. Sci.*, vol. 38, no. 1, pp. 38–75, 2014.

[46] B. Rottman, J. Kominsky, and F. Keil, "Children use temporal cues to learn causal directionality," *Cogn. Sci.*, vol. 38, no. 3, pp. 489–513, 2014.

[47] B. Rottman and F. Keil, "Learning causal direction from repeated observations over time," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, 2011.

[48] P. Gördenfors, "Events and causal mappings modeled in conceptual spaces," *Front. Psychol.*, vol. 11, p. 630, 2020.

[49] P. Gördenfors and M. Lombard, "Causal cognition, force dynamics and early hunting technologies," *Front. Psychol.*, vol. 9, p. 87, 2018.

[50] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[51] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097-1105.

[52] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *Intelligent Systems, IEEE*, vol. 24, pp. 8–12, 2009.

[53] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners for task-oriented dialogue systems," arXiv:2008.06239, 2020.

[54] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," in *WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018.

[55] S. K. Sgaier, V. Huang, and G. Charles, “The case for causal AI,” *Stanf. Soc. Innov. Rev.*, vol. 18, pp. 50–55, 2020.

[56] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, *Invariant risk minimization*, arXiv:1907.02893, 2019.

[57] T. Hellström, V. Dignum, and S. Bensch, “Bias in machine learning – What is it good for?,” in *Proceedings of 1st International Workshop on New Foundations for Human-Centered AI* (*NeHuAI*) *at ECAI-2020*, 2020.

[58] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, et al., “Toward causal representation learning,” *Proceedings of the IEEE*, 2021, pp. 1–23.

[59] R. Guo, L. Cheng, J. Li, P. Hahn, and H. Liu, “A survey of learning causality with data – problems and methods,” *ACM Computing Surveys* (*CSUR*), vol. 53, pp. 1–37, 2018.

[60] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, et al., “A meta-transfer objective for learning to disentangle causal mechanisms,” arXiv:1901.10912 [cs.LG], 2019.

[61] O. Goudet, D. Kalainathan, P. Caillou, D. Lopez-Paz, I. Guyon, and M. Sebag, “Learning functional causal models with generative neural networks,” in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, H. Escalante, et al., Eds., The Springer Series on Challenges in Machine Learning. Springer, Cham, 2018, pp. 39–80.

[62] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. Balasubramanian, “Neural network attributions: A causal perspective,” in *Proceedings of the 36th International Conference on Machine Learning* (*ICML*), 2019.

[63] F. Johansson, U. Shalit, and D. Sontag, “Learning representations for counterfactual inference,” in *Proceedings of Machine Learning Research*, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48, New York, USA: PMLR, 20–22 Jun 2016, pp. 3020–3029.

[64] J. Peters, P. Bühlmann, and N. Meinshausen, “Causal inference by using invariant prediction: identification and confidence intervals,” *J. R. Stat. Soc. Series B* (*Stat Methodol*)., vol. 78, no. 5, pp. 947–1012, 2016.

[65] C. Hofsten, “Eye–hand coordination in newborns,” *Dev. Psychol.*, vol. 18, pp. 450–461, 1982.

[66] D. Caligiore, T. Ferrauto, D. Parisi, N. Accornero, M. Capozza, and G. Baldassarre, “Using motor babbling and Hebb rules for modeling the development of reaching with obstacles and grasping,” in *Proceedings of International Conference on Cognitive Systems* (*COGSYS 2008*), 2008.

[67] Z. Mahoor, B. J. MacLennan, and A. C. McBride, “Neurally plausible motor babbling in robot reaching,” in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics* (*ICDL-EpiRob*), 2016, pp. 9–14.

[68] G. Schillaci and V. Hafner, “Prerequisites for intuitive interaction – on the example of humanoid motor babbling,” in *Proceedings of the Workshop on The Role of Expectations in Intuitive Human-Robot Interaction* (*at HRI 2011*), 2011, pp. 23–27.

[69] Y. Demiris and A. Dearden, *From motor babbling to hierarchical learning by imitation: a robot developmental pathway*, 2005, http://cogprints.org/4961/.

[70] Z. Liu, A. Hitzmann, S. Ikemoto, S. Stark, J. Peters, and K. Hosoda, “Local online motor babbling: Learning motor abundance of a musculoskeletal robot arm*,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), 2019, pp. 6594–6601.

[71] P. Lanillos, E. Dean-León, and G. Cheng, “Yielding self-perception in robots through sensorimotor contingencies,” *IEEE Trans. Cogn. Develop. Syst.*, vol. 9, 2017, pp. 100–112.

[72] J. Brawer, M. Qin, and B. Scassellati, “A causal approach to tool affordance learning,” in *Intelligent Robots and Systems* (*IROS 2020*), 2020.

[73] A. A. Bhat and V. Mohan, “Causal learning by a robot with semantic-episodic memory in an Aesopas fable experiment,” arXiv:2003.00274, 2020.

[74] A. A. Bhat, V. Mohan, G. Sandini, and P. Morasso, “Humanoid infers Archimedes’ principle: understanding physical relations and object affordances through cumulative learning experiences,” *J. R. Soc. Interface*, vol. 13, 20160310, 2016.

[75] K. Dickerson, P. Gerhardstein, and A. Moser, “The role of the human mirror neuron system in supporting communication in a digital world,” *Front. Psychol.*, vol. 8, p. 698, 2017.

[76] M. Longo, A. Kosobud, and B. Bertenthal, “Automatic imitation of biomechanically possible and impossible actions: effects of priming movements versus goals,” *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 34, no. 2, pp. 489–501, 2008.

[77] J. Decety and P. L. Jackson, “The functional architecture of human empathy,” *Behav. Cogn. Neurosci. Rev.*, vol. 3, pp. 71–100, 2004.

[78] A. Meltzoff and J. Decety, “What imitation tells us about social cognition: A rapprochement between developmental psychology and cognitive neuroscience,” *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 358, pp. 491–500, 2003.

[79] G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga, “Understanding mirror neurons: A bio-robotic approach,” *Interact. Stud.*, vol. 7, pp. 197–232, 2006.

[80] K. Rebrová, M. Pecháč, and I. Farkaš, “Towards a robotic model of the mirror neuron system,” in *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics* (*ICDL*), 2013, pp. 1–6.

[81] S. Thill, H. Svensson, and T. Ziemke, “Modeling the development of goal-specificity in mirror neurons,” *Cogn. Comput.*, vol. 3, pp. 525–538, 2011.

[82] M. Marraffa, “Theory of mind,” in *The Internet Encyclopedia of Philosophy*, J. Fieser and B. Dowden, Eds., 2020. https://www.iep.utm.edu/, March 31, 2021.

[83] Y. Barnes-Holmes, L. McHugh, and D. Barnes-Holmes, “Perspective-taking and theory of mind: A relational frame account,” *Behav. Anal. Today*, vol. 5, pp. 15–25, 2004.

[84] M. Persiani and T. Hellström, “Intent recognition from speech and plan recognition,” in *Advances in Practical Applications of Agents, Multi-Agent Systems, and Trustworthiness, The PAAMS Collection 2020, Lecture Notes in Computer Science, vol. 12092*, Y. Demazeau, T. Holvoet, J. Corchado, and S. Costantini, Eds., Cham, Switzerland: Springer, 2020, pp. 212–223.

[85] C. Baker and J. Tenenbaum, “Modeling human plan recognition using bayesian theory of mind,” in *Plan, Activity, and Intent Recognition: Theory and Practice*, G. Sukthankar, C. Geib, H. H. Bui, D. V. Pynadath, and R. P. Goldman, Eds., San Francisco: Morgan Kaufmann, 2014, pp. 177–204.

[86] S. Levine and B. Williams, “Concurrent plan recognition and execution for human-robot teams,” in *Proceedings*

International Conference on Automated Planning and Scheduling, ICAPS, vol. 2014, 2014, pp. 490–498.

[87] E. A. Billing and T. Hellström, "A formalism for learning from demonstration," *Paladyn, J. Behav. Robot.*, vol. 1, no. 1, pp. 1–13, 2010.

[88] E. A. Billing, T. Hellström, and L.-E. Janlert, "Behavior recognition for learning from demonstration," in *Proceedings of IEEE International Conference on Robotics and Automation* (*ICRA 2010*), (Anchorage, Alaska), 2010, pp. 866–872.

[89] A. Hussein, M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Comput. Surv.*, vol. 50, 2017.

[90] D. Angelov and S. Ramamoorthy, "Learning from demonstration of trajectory preferences through causal modeling and inference," in *Robotics Science and Systems* (*RSS*) *Workshop*, 2018.

[91] D. Angelov, Y. Hristov, and S. Ramamoorthy, "Using causal analysis to learn specifications from task demonstrations," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, (Richland, SC), International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1341–1349.

[92] K. Noda, H. Arie, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robot. Auton. Syst.*, vol. 62, 2014.

[93] G. E. Katz, D.-W. Huang, R. Gentili, and J. Reggia, "Imitation learning as cause-effect reasoning," in *Proceedings of the 9th Conference on Artificial General Intelligence*, New York, USA: Springer International Publishing, 2016.

[94] G. Katz, D. Huang, T. Hauge, R. Gentili, and J. Reggia, "A novel parsimonious cause-effect reasoning algorithm for robot imitation and plan recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 10, no. 2, pp. 177–193, 2018.

[95] J. Zhang, D. Kumor, and E. Bareinboim, "Causal imitation learning with unobserved confounders," Tech. Rep. R-66, University of Columbia, CausalAI Laboratory, 2020.

[96] H. B. Suay, J. Beck, and S. Chernova, "Using causal models for learning from demonstration," in *AAAI Fall Symposium: Robots Learning Interactively from Human Teachers*, 2012.

[97] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson, "The tetrad project: Constraint based aids to causal model specification," *Multivariate Behav. Res.*, vol. 33, 2002.

[98] P. D. Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 11693–11704.

[99] C. Xiong, N. Shukla, W. Xiong, and S.-C. Zhu, "Robot learning with a spatial, temporal, and causal and-or graph," in *2016 IEEEE International Conference on Robotics and Automation* (*ICRA*), 2016, pp. 2144–2151.

[100] Y. Li, D. Zhang, F. Yin, and Y. Zhang, "Operation mode decision of indoor cleaning robot based on causal reasoning and attribute learning," *IEEE Access*, vol. 8, pp. 173376–173386, 2020.

[101] B. Meadows, M. Sridharan, and Z. Colaco, "Towards an explanation generation system for robots: Analysis and recommendations," *Robotics*, vol. 5, p. 21, 2016.

[102] M. Beetz and H. Grosskreutz, "Causal models of mobile service robot behavior," in *Proceedings of the Fourth International Conference on AI Planning Systems* (*AIPS*), 1998.

[103] P. R. Cohen, C. Sutton, and B. Burns, "Learning effects of robot actions using temporal associations," in *Proceedings of the 2nd International Conference on Development and Learning, ICDL 2002*, 2002, pp. 96–101.

[104] N. Shukla, C. Xiong, and S. Zhu, "A unified framework for human-robot knowledge transfer," in *AAAI Fall Symposia*, 2015, pp. 125–127.

[105] E. Aker, A. Erdogan, E. Erdem, and V. Patoglu, "Causal reasoning for planning and coordination of multiple housekeeping robots," in *Proceedings of the 11th International Conference on Logic Programming and Nonmonotonic Reasoning*, LPNMR'11, Berlin, Heidelberg: Springer-Verlag, 2011, pp. 311–316.

[106] E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, and H. Turner, "Nonmonotonic causal theories," *Artif. Intell.*, vol. 153, no. 1–2, pp. 49–104, 2004.

[107] N. Mccain and H. Turner, "Causal theories of action and change," in *Proceedings of the AAAI-97*, 1997.

[108] J. Ji and X. Chen, "From structured task instructions to robot task plans," in *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (*KEOD-2013*), 2013, pp. 237–244.

[109] S. C. Smith and S. Ramamoorthy, "Counterfactual explanation and causal inference in service of robustness in robot control," arXiv:2009.08856, 2020.

[110] D. Lagnado, "Causal thinking," in *Causality in the Sciences*, P. McKay Illari, F. Russo, and J. Williamson, Eds., United Kingdom: Oxford University Press, 2011, pp. 129–149.

[111] M. Haidle, M. Bolus, M. Collard, N. Conard, D. Garofoli, M. Lombard, et al., "The nature of culture: an eight-grade model for the evolution and expansion of cultural capacities in hominins and other animals," *J. Anthropol. Sci.*, vol. 93, pp. 43–70, 2015.

[112] T. Hellström and S. Bensch, "Understandable robots," *Paladyn, J. Behav. Robot.*, vol. 9, pp. 110–123, 2018.

[113] T. Lombrozo and N. Vasilyeva, "Causal explanation," in *The Oxford Handbook of Causal Reasoning*, M. R. Waldmann, Ed., Oxford, UK: Oxford University Press, 2017, pp. 415–432.

[114] B. Skow, "Are there non-causal explanations (of particular events)?," *Br. J. Philos. Sci.*, vol. 65, pp. 445–467, 2013.

[115] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, et al., "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Sci. Robot.*, vol. 4, no. 37, eaay4663, 2019.

[116] M. Sridharan and B. Meadows, "Towards a theory of explanations for human-robot collaboration," *Künstliche Intell.*, vol. 33, no. 4, pp. 331–342, 2019.

[117] M. Zheng and S. Kleinberg, "A method for automating token causal explanation and discovery," in *FLAIRS Conference*, 2017.

[118] E. Erdem, E. Aker, and V. Patoglu, "Answer set programming for collaborative housekeeping robotics: Representation, reasoning, and execution," *Intell. Serv. Robot.*, vol. 5, pp. 275–291, 2012.

[119] E. Erdem, K. Haspalamutgil, V. Patoglu, and T. Uras, "Causality-based planning and diagnostic reasoning for cognitive factories," in *IEEE 17th International Conference on*

Emerging Technologies Factory Automation (*ETFA 2012*), 2012, pp. 1–8.

[120] A. Singh, N. Baranwal, K.-F. Richter, T. Hellström, and S. Bensch, "Towards verbal explanations by collaborating robot teams," in *1st Workshop on Quality of Interaction in Socially Assistive Robots, ICSRa19*, Madrid, Spain, 2019.

[121] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, "Causal reasoning for algorithmic fairness," *CoRR*, vol. abs/1805.05859, 2018. Available at: http://arxiv.org/abs/1805.05859.

[122] G. DodigCrnkovic and D. Persson, "Sharing moral responsibility with robots: A pragmatic approach," in *Proceedings of the 2008 Conference on Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008*, IOS Press, 2008, pp. 165–168.

[123] M. Moore, "Causation in the law," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Metaphysics Research Lab, Stanford University, winter 2019 ed., 2019.

[124] T. Hellström, "On the moral responsibility of military robots," *Ethics Inf. Technol.*, vol. 15, pp. 99–107, 2013.

[125] R. Hakli and P. Makela, "Moral responsibility of robots and hybrid agents," *Monist*, vol. 102, pp. 259–275, 2019.

[126] A. Sharkey, "Can robots be responsible moral agents? And why should we care?," *Connect. Sci.*, vol. 29, pp. 210–216, 2017.

[127] J. Kober, J. Bagnell, and J. Peters, "Reinforcement learning in robotics: a survey," *Int. J. Robot. Res.*, vol. 32, pp. 1238–1274, 2013.

[128] S. J. Gershman, "Reinforcement learning and causal models," in *The Oxford Handbook of Causal Reasoning*, M. R. Waldmann, Ed., Oxford, UK: Oxford University Press, 2017, pp. 295–306.

[129] J. Zhang and E. Bareinboim, "Designing optimal dynamic treatment regimes: A causal reinforcement learning approach," Tech. Rep. R-57, University of Columbia, CausalAI Laboratory, 2020.

[130] G. Leroy, *Designing User Studies in Informatics*, London: Springer-Verlag, 2011.

[131] R. E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*, 4th ed., Thousand Oaks, CA: SAGE Publishing, 2013.

[132] T. Haavelmo, "The statistical implications of a system of simultaneous equations," *Econometrica*, vol. 11, no. 1, pp. 1–12, 1943.

[133] Y. Yamashita, H. Ishihara, T. Ikeda, and M. Asada, "Appearance of a robot influences causal relationship between touch sensation and the personality impression," in *Proceedings of the 5th International Conference on Human Agent Interaction* (*HAI '17*), 2017, pp. 457–461.

[134] D. Dash, M. Voortman, and M. Jongh, "Sequences of mechanisms for causal reasoning in artificial intelligence," in *IJCAI International Joint Conference on Artificial Intelligence*, 2013, pp. 839–845.

[135] L. Mireles-Flores, "Recent trends in economic methodology: a literature review," *Research in the History of Economic Thought and Methodology*, vol. 36, no. A, pp. 93–126, 2018.

[136] P. Fredriksson and M. Söderström, "The equilibrium impact of unemployment insurance on unemployment: Evidence from a non-linear policy rule," *J. Pub. Econom.*, vol. 187, 104199, 2020.

[137] J. Y. Halpern, "Appropriate causal models and the stability of causation," *Rev. Symbol. Logic*, vol. 9, no. 1, pp. 76–102, 2016.